Rethinking the Force Concept Inventory: Developing a Cognitive Diagnostic Assessment to Measure Misconceptions in Newton's Laws

Mary Armistead Norris

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Educational Research and Evaluation

Gary Skaggs, Chair
George Glasson
David Kniola
Yasuo Miyazaki

September 17, 2021
Blacksburg, Virginia

Keywords: cognitive diagnostic modeling, physics education, DINA, psychometrics, misconceptions, concept inventory

Rethinking the Force Concept Inventory: Developing a Cognitive Diagnostic Assessment to Measure Misconceptions in Newton's Laws

Mary Armistead Norris

ABSTRACT

Student misconceptions in science are common and may be present even for students who are academically successful. Concept inventories, multiple-choice tests in which the distractors map onto common, previously identified misconceptions, are commonly used by researchers and educators to gauge the prevalence of student misconceptions in science. Distractor analysis of concept inventory responses could be used to create profiles of individual student misconceptions which could provide deeper insight into the phenomenon and provide useful information for instructional planning, but this is rarely done as the inventories are not designed to facilitate it. Researchers in educational measurement have suggested that diagnostic cognitive models (DCMs) could be used to diagnose misconceptions and to create such misconception profiles. DCMs are multidimensional, confirmatory latent class models which are designed to measure the mastery/presence of fine-grained skills/attributes. By replacing the skills/attributes in the model with common misconceptions, DCMs could be used to filter students into misconception profiles based on their responses to concept inventory-like questions. A few researchers have developed new DCMs that are specifically designed to do this and have retrofitted data from existing concept inventories to them. However, cognitive diagnostic assessments, which are likely to display better model fit with DCMs, have not been developed. This project developed a cognitive diagnostic assessment to measure knowledge and misconceptions about Newton's laws and fitted it with the deterministic input noisy-and-gate (DINA) model. Experienced physics instructors assessed content validity and Q-matrix alignment. A pilot test with 100 undergraduates was conducted to assess item quality within a classical test theory framework. The final version of the assessment was field tested with 349 undergraduates. Results showed that response data displayed acceptable fit to the DINA model at the item level, but more questionable fit at the overall model level; that responses to selected items were similar to those given to two items from the Force Concept Inventory; and that, although all students were likely to have misconceptions, those with lower knowledge scores were more likely to have misconceptions.

Rethinking the Force Concept Inventory: Developing a Cognitive Diagnostic Assessment to Measure Misconceptions in Newton's Laws

Mary Armistead Norris

GENERAL AUDIENCE ABSTRACT

Misconceptions about science are common even among well-educated adults. Misconceptions range from incorrect facts to personal explanations for natural phenomena that make intuitive sense but are incorrect. Frequently, they exist in people's minds alongside correct science knowledge. Because of this, misconceptions are often difficult to identify and to change. Students may be academically successful and still retain their misconceptions. Concept inventories, multiple-choice tests in which the incorrect answer choices appeal to students with common misconceptions, are frequently used by researchers and educators to gauge the prevalence of student misconceptions in science. Analysis of incorrect answer choices to concept inventory questions can be used to determine individual student's misconceptions, but it is rarely done because the inventories are not known to be valid measures for this purpose. One source of validity for tests is the statistical model that is used to calculate test scores. In valid tests, student's answers to the questions should follow similar patterns to those predicted by the model. For instance, students are likely to get questions about the same things either all correct or all incorrect. Researchers in educational measurement have proposed that certain types of innovative statistical models could be used to develop tests that identify student's misconceptions, but no one has done so. This project developed a test to measure knowledge and misconceptions about forces and assessed how well it predicted student's misconceptions compared to two statistical models. Results showed that the test predicted student's knowledge in good agreement and misconceptions in moderate agreement with the statistical models; that students tended to answer selected questions in the same way that they answered two similar questions from an existing test about forces; and that, although students with lower test scores were more likely to have misconceptions, students with high test scores also had misconceptions.

Acknowledgements

My deepest gratitude goes to my advisor Gary Skaggs for his unwavering support and for introducing me to the field of educational measurement. This project was built upon a foundation of your kindness, patience, knowledge, and critical eye. Thank you for sharing them with me.

I am also deeply indebted to George Glasson, David Kniola, and Yasuo Miyazaki for their interest in my project and their invaluable advice and suggestions. The input from each of you has made my work stronger. Thank you for serving on my committee.

I am extremely grateful to Nancy Bodenhorn for generously supporting my studies. You introduced me to the world of accreditation and modeled how to get things done in academia while always putting people first. Thank you for taking care of me.

I am also grateful to Gerard Lawson and Laura Welfare for showing me how a research team operates and providing me the opportunity to wrestle with big, messy data sets. I use these skills in my work almost every day.

Thanks also goes to Elizabeth Creamer who allowed me to enroll as a part-time student and assured me that I would not be too old to start a new career by the time I finished. Without your encouragement, I would never have begun.

I could not have done this without the support of my family.

Thanks to my parents. Although you will never know that I finished, I know that you would be proud of me if you were here.

Thank you to my sisters Patty Norris and Priscilla Masters. I don't need to tell you why. We are sisters.

Thank you to my children Nick Britten, Peter Britten, and Asa Britten for their constant belief in my abilities and for filling my heart with love every day. Your hugs and kind words always keep me going.

Finally, thanks to my husband, Dan Britten for staying by my side, for tending to the daily feeding and watering of our family so that I could learn and grow, and for reminding me to take care of my body and spirit as well as my mind. You are my wings and my heart.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

1-PL, 2-PL, 3-PL	One-parameter logistic, Two-parameter logistic, Three parameter logistic
AIC	Akaike Information Criterion
AP	Advanced Placement
BIC	Bayesian Information Criterion
Bug-DINO	Bug-diagnostic noisy-or-gate
CDA	Cognitive Diagnostic Assessment
CFA	Confirmatory Factor Analysis
CTT	Classical Test Theory
DCM	Diagnostic Cognitive Model
DINA	Deterministic input noisy-and-gate
EAP	Expected a posteriori
EFA	Exploratory Factor Analysis
EM	Expectation maximization
FCI	Force Concept Inventory
IB	International Baccalaureate
ICC	Item Characteristic Curve
IRT	Item Response Theory
JMLE	Joint maximum likelihood estimation
LCA	Latent Class Analysis
MAFA	Misconceptions About Force Assessment
MAP	Maximum a posteriori
MMLE	Marginal maximum likelihood estimation
NC-RUM	Noncompensatory reparameterized unified model
NIDA	Noisy deterministic-and-gate
RMSEA	Root mean square error of approximation
SEM	Standard error of measurement
SICM	Scaling Individuals and Classifying Misconceptions

SISM	Simultaneously Identifying Skills and Misconceptions
SRMSR	Standardized root mean square root of squared residuals
STEM	Science, technology, engineering, and mathematics

Chapter 1

Introduction to the Study

In the US there has long been a focus on improving science, technology, engineering, and mathematics (STEM) education. Over the past 60 years as the economy has become more global, the focus of STEM education in the US has broadened from training more highly skilled STEM workers (Government Publishing Office [GPO], 1958; Powell, 2007) to increasing the quality of STEM education for all children (National Research Council [NRC], 2012). In 1990, the American Association for the Advancement of Science published *Science for All Americans*. The authors emphasized the need for a scientifically literate citizenry:

> What the future holds in store for individual human beings, the nation, and the world depends largely on the wisdom with which humans use science and technology. And that, in turn, depends on the character, distribution, and effectiveness of the education that people receive. (Rutherford & Ahlgren, 1990, p. *xiii*)

Yet despite the continued and increasing influence of scientific and technological advances on our lives, many American workers—even recent graduates—lack fundamental STEM skills and knowledge and the call for more effective STEM education continues (National Science Foundation [NSF], 2020).

Physics describes how matter and energy act. Because matter and energy are the main building blocks of the universe, physics is the cornerstone of the other basic sciences and understanding physics is of special importance to understanding the reasons behind what is learned in other sciences (Feynman et al., 1964). There is growing concern that students who do

not understand science, especially physics, may have difficulty understanding the other basic sciences and engaging in an increasingly technological world (Bessin, 2007; Feierman et al., 2006; National Academy of Sciences et al., 2007; National Commission on Excellence in Education, 1983; NRC, 2012; NSF, 2020; White, 2008). Citizens who lack an understanding of basic sciences may not be able to follow the reasoning of scientific discoveries or make informed judgements about current issues such as climate change or nanoscience. The primary concern is to create an educated citizenry that is prepared to address important social and scientific issues such as climate change (Meltzer et al., 2012; White, 2008). An accurate understanding of basic physics is a necessary building block and an important indicator of a person's scientific literacy.

Only 39% of US high school students have completed at least one course in physics at graduation (Meltzer et al., 2012), and even this minority of students may not be as knowledgeable in physics as their international counterparts (Provasnik et al., 2016). Secondary physics education in the US is plagued by multiple challenges. Courses tend to be less rigorous than in many other countries (Provasnik et al., 2016) and, due to a shortage of physics teachers, many courses are taught by teachers who lack an appropriate physics background (Meltzer et al., 2012). Compounding the problem is the fact that students who complete a traditional physics course tend to retain many misconceptions about the physical world (Halloun & Hestenes, 1985a). The prevalence of misconceptions among science students is well known. While existing instruments are useful for measuring the prevalence of misconceptions as a whole, they do not provide psychometrically sound profiles of individual student's misconceptions.

My research investigated the creation of a tool to identify students' physics misconceptions so that the misconceptions can be addressed through instruction. The purpose of this study was to design and validate an instrument that would serve as a diagnostic tool for

introductory physics students. The instrument that I developed is the *Misconceptions About Force Assessment (MAFA)*. The MAFA measures knowledge of Newton's first and second laws of motion and diagnoses the possession of six common misconceptions about the laws. In addition to developing and validating the instrument, I used scores from the instrument to investigate the relationship between students' knowledge of Newton's first and second laws and their possession of the six misconceptions.

In the following sections, I will provide a brief history of US focus on STEM education, an outline of how US students perform compared to students in other nations, a description of the current state of K-12 physics education in the US, and an example of how educational measurement can be applied to strengthen STEM education. This will be followed by the specific aims of this study through its purpose statement and research questions. I also provide significance of this study to future research, policy, and practice as well as known delimitations to this study.

STEM Education in the US

Historical Context

The origin of the US focus on improvement of K-12 STEM education can be traced to the launch of Sputnik in 1957. After the launch, there was a fear that the US was falling behind the USSR in terms of technology development. Congress reacted by passing the National Defense Education Act (NDEA) one purpose of which was "to encourage and assist in the expansion and improvement of educational programs to meet critical national needs" (GPO, 1958, p.1580). The Act provided funding for math, science, and foreign language education; for research into how to better use technology in education; for vocational training for a stronger workforce; and to

implement testing to identify gifted students (GPO, 1958; Powell, 2007). It also doubled the funding for the National Science Foundation (NSF), an organization which supports the development of science curricula and training of science teachers as well as research (Hechinger Report, 2011). The primary focus of the Act was to develop highly skilled STEM workers.

Later calls for improving STEM education continued to address the need for highly skilled workers, but also emphasized the need for all citizens to possess a solid STEM education. In 1983, the National Commission on Excellence in Education (NCEE) issued *A Nation at Risk: The Imperative for Educational Reform*. According to the report, post-Sputnik gains in science and engineering had been lost at the same time that globalization had increased the demand for more highly skilled STEM workers and a more complex world had increased the need for STEM knowledge for all. The position was that citizens who lacked STEM knowledge would be barred from full participation in our democratic society, and that, while the average citizen was better educated than a generation ago, the average high school or college graduate was not as well educated. Among the recommendations of the Commission, were government support for the development of more challenging math and science standards and recruiting more highly qualified math and science teachers. Just over two decades later, the National Academy of Sciences, National Academy of Engineering, and Institute of Medicine released *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (2007). The authors highlight many of the same problems in the education system including a paucity of highly qualified math and science teachers and a system that does not adequately prepare students with "the interest, motivation, knowledge, and skills they will need to compete and prosper in the emerging world" (p. 94). As will be explained in a later section of this chapter,

physics education is key to understanding the other sciences and should be a focus in efforts to improve STEM education.

International Comparisons

Despite the actions taken to improve K-12 STEM education in the US, international comparisons of student achievement show that US secondary school students have average performance at best (Organisation for Economic Cooperation and Development [OECD], 2018). The US has participated in every administration of the most popular international measure of secondary-level math and science achievement, the Programme for International Student Assessment (PISA). The Programme assesses knowledge and skills in science, mathematics, and reading among students at age 15. Items on the assessment are designed to measure how well students can apply knowledge and skills they have learned in school—a design which reflects industry needs for an appropriately prepared workforce (OECD, 2018). In 2015, seventy-two countries participated in the assessment (OECD, 2018). Results showed the US had below average performance in mathematics and average performance in both reading and science—a rating that has been relatively stable, but arguably unacceptable, over the course of the Programme (OECD, 2016).

Another measure, the Trends in International Mathematics and Science Study (TIMSS) Advanced, is less widely used but provides additional evidence. This assessment measures math and physics achievement among students in their final year of secondary school. In 2015, nine countries participated in the physics portion and the US scored about average—scoring higher than three countries, lower than four, and about equal to one other country (Provasnik et al., 2016). What is more disturbing than average performance, however, is that US physics courses are less demanding than those in the other countries (Provasnik et al., 2016). TIMSS Advanced

reports the coverage rate—the percent of 18-year-olds who have either taken or are enrolled in a physics course that covers a set list of topics--for each participating education system. In 2015, the only types of US courses that adequately covered the topics were AP, IB, or second-year physics courses. This put the US coverage rate at 4.8%--second only to Lebanon. In comparison, France, Italy, and Slovenia have coverage rates of 21.5%, 18.2%, and 14.3% respectively (Provasnik et al., 2016). Even though the number of US high school students taking an AP or second course in physics has increased ninefold in the last two decades (Meltzer et al., 2012), it is clear that the United States lags behind other developed nations.

Why Focus on Physics?

Physics is often described as the most fundamental science discipline (Feynman et al., 1964). The laws of physics describe the interactions of matter and energy of which everything in the known universe is composed and it is a foundational science that underpins the fundamental processes of chemistry, biology, and earth science. Because it is key to understanding other sciences, it becomes imperative that all students have the opportunity to complete at least one course in physics (Bessin, 2007; Feierman et al., 2006; White, 2008). Historically, US high school students have taken physics as a final science course. Given this course sequence, only 30% of US students complete a physics course in high school (Feierman et al., 2006). The *Physics First* movement works to increase student enrollment in physics by inverting the typical sequence of science courses taken in high school (Ewald et al., 2005). *Physics First* suggests that students complete physics before biology or chemistry and that topics covered in Earth Science courses be woven into the other three. According to a survey of physics teachers, students in schools that follow this model take more science courses (Feierman et al., 2006). Currently, 39%

of high school graduates have taken at least one course in physics—a number which has doubled in the last two decades (Meltzer et al., 2012).

The best predictor of student achievement in STEM courses is having a teacher who is certified and has a degree in the field (National Academy of Sciences et al., 2007). Even though fewer than half of high school students take physics, the supply of physics teachers with a degree in the field has not kept up with physics enrollments. According to a report from the Task Force on Teacher Education in Physics (T-TEP), there is a severe, long-term shortage of qualified physics teachers in the US (Meltzer et al., 2012). This shortage poses a great challenge to offering students who enroll in physics a quality physics education. As increasing numbers of students take physics in high school (twice as many complete one course and nine times as many take an AP or second course as 20 years ago) the problem has been exacerbated (Meltzer et al., 2012). The T-TEP report states that 39% of recent high school graduates have taken at least one course in physics, but that only 47% of those courses are taught by a teacher with either a physics or physics education degree. For comparison, 73% of biology courses and 80% of humanities courses are taught by an educator with a degree in the field (Meltzer et al., 2012).

Most physics teachers did not receive training in physics pedagogy and many received little formal physics education; they consequently "develop their skills through on-the-job practice, teaching a subject that they never intended to teach, nor were trained to teach" (Meltzer et al., 2012, p.13). The large number of students who complete a physics course and continue to have misconceptions about the physical world may be impacted by this lack of preparation in the teaching force. Students tend to enter physics with well-established ideas about how the world acts, but many of these ideas are incorrect (Halloun & Hestenes, 1985a). These incorrect ideas hinder students' progress in learning physics and are resistant to change (Halloun & Hestenes,

1985a). High achievement test scores are not always indicators of conceptual understanding—they may occur despite incorrect structuring of ideas or ideas themselves (Brown & Hammer, 2013; Harrison & Treagust, 2001). Even successful physics students may retain misconceptions upon completion of a physics course. Identifying student misconceptions in physics can be difficult. It is easy to mistake rote learning for deeper understanding. Physics teachers who lack a strong physics background may have difficulty recognizing and addressing students' misconceptions. Even experienced teachers (who are more likely to recognize misconceptions) may be limited in their ability to diagnose individual students' misconceptions because of the large number of students they teach. Given that one of the most valuable practices for learning is to provide feedback to students about their misconceptions along with opportunities to correct them (Hattie, 2015), an assessment instrument which provides feedback on student misconceptions may provide valuable diagnostic information for all levels of physics teachers and physics students.

Currently, there are many concept inventories—multiple choice assessments which measure a set of core knowledge—in physics. Commonly used inventories such as the Force Concept Inventory (FCI) (Hestenes et al., 1992) and the Force and Motion Conceptual Evaluation (Thornton & Sokoloff, 1998) use misconceptions as distractors. However, none of these inventories are typically scored to provide information about individual students' specific misconceptions. Scores are calculated by summing correct answers to provide a total score and incorrect answers—those which might indicate the presence of a misconception—are not scored. Students with higher scores are presumed to have fewer misconceptions. Alternate scoring of the FCI to diagnose misconceptions at the classroom level has been explored and the possibility that the test is multidimensional has been proposed. This research will be summarized in Chapter 2.

How Educational Measurement Can Help

Despite the importance of misconceptions, none of the existing instruments provide a reliable and efficient method for diagnosing individual students' misconceptions. In fact, most concept inventories, like most large-scale tests, are based on one of two unidimensional psychometric models, classical test theory (CTT) or item response theory (IRT). Both provide estimates of a single individual ability, such as physics knowledge, as measured on a continuous scale. Identifying multiple misconceptions requires the use of more complex multidimensional measurement models such as multidimensional item response theory (MIRT) or a diagnostic cognitive model (DCM). Diagnostic cognitive models have been "developed specifically for the purpose of identifying the presence or absence of multiple fine-grained skills" (de la Torre, 2009, p.163). Instead of locating an individual within a group of respondents as in CTT or along a continuous ability scale as in IRT, DCMs produce an individual profile which indicates skills that have and have not been mastered.

Psychometricians have developed numerous DCMs, but few cognitive diagnostic assessments (CDAs) based on them. DCMs in which skills are replaced with misconceptions provide the potential to measure student misconceptions (Bradshaw & Templin, 2014; de la Torre, 2009; DiBello et al., 2015). Recently, researchers have developed a small number of DCMs in which skills are replaced with misconceptions, but no CDAs based on them (Bradshaw & Templin, 2014; Kuo et al., 2018; Kuo et al., 2016). Because there are no assessments which have been designed to fit these models, data from existing assessments including the FCI have been retrofitted to them. Although a common practice due to the lack of CDAs, retrofitting data that were designed for a unidimensional model has been shown to result in examinee misclassification and poor model and item fit (de la Torre & Minchen, 2014; Lee et al., 2012;

Rupp & Templin, 2008). This lack of CDAs is likely compounded by the need for expertise in both psychometrics and subject-matter knowledge which are required for their development (de la Torre, 2009). As a subject for CDA development, Newtonian physics provides a well-documented set of misconceptions and STEM education provides a need for them to be measured.

About the Study

Problem Statement

In educational measurement, the creation of statistical models has greatly outpaced the application of models to create model-based assessments. Researchers have demonstrated the feasibility of using DCMs to measure misconceptions using simulated or retrofitted data (Bradshaw & Templin, 2014; Kuo et al., 2018) but have not developed new assessments based on such models. Retrofitted data come from concept inventories in which distractors are expressions of common misconceptions.

An example of this is the Force Concept Inventory (FCI), a conceptual test that is widely used in introductory physics courses to gauge student mastery of Newtonian thinking. Each item has five options and many of the distractors have been mapped onto common student misconceptions in the topics (Hestenes et al., 1992). A few studies have suggested methods to use incorrect answers to measure misconceptions (Bao & Reddish, 2001; Fulmer, 2015; Martin-Blas et al., 2010; Saivinainen & Scott, 2002a, 2002b; Savinainen & Viiri, 2008; Yasuda & Taniguchi, 2013), but none of these methods results in a psychometrically sound profile of misconceptions for individuals. The FCI is most often used simply to measure student growth in understanding by comparing pre- and post-instruction total scores. A more sophisticated measure

is needed that specifically addresses misconceptions. One way to accomplish this is through a test development process that uses advanced psychometric models such as DCMs particularly in educational settings (DiBello et al., 2007). The application of a DCM to create a student-level assessment to measure misconceptions addresses this need.

Research on conceptual change shows that the process of learning physics is messy. Students do not simply replace existing misconceptions with correct physics knowledge. New information must be incorporated into an existing framework of ideas and beliefs, some of which contradict scientific thinking (Posner et al., 1982). Instruction often facilitates the development of misconceptions as students distort the scientific information to fit their existing knowledge (Vosniadou & Skopeliti, 2014). The result is a mixture of correct and incorrect scientific knowledge. In other words, students may solve problems correctly even when they have misconceptions (Posner & Gertzog, 1982). Current DCMs may underidentify misconceptions because they cannot diagnose specific misconceptions that coexist with correct knowledge (Bradshaw & Templin, 2014; Kuo et al., 2018). A CDA based on a DCM that can identify coexisting knowledge and specific misconceptions would fill a valuable role in science education.

DCMs require the specification of a Q-matrix, a matrix showing what skills and knowledge are needed to answer each item correctly. Entries in the matrix are a "1" for each skill that is needed to choose a particular answer and a "0" for each that is not required. Teams of content experts specify which skills are required to choose each answer. An accurate Q-matrix is essential for a well-estimated model (de la Torre, 2008; Liu et al., 2012). As the size and complexity of the Q-matrix increase, the potential for misspecification of the matrix also increases and the model may be poorly estimated (de la Torre, 2008; Liu et al., 2012). Larger and

more complex Q-matrices also require longer assessments, more respondents, and more time to estimate well. To keep CDAs at a reasonable length, "it is necessary to keep models as simple as possible while satisfying the constraints imposed by the diagnostic purpose" (DiBello et al., 2007, p.985). In a traditional DCM, only correct answer choices are scored. Therefore, the Q-matrix contains one row for each item. Two existing DCMs for measuring misconceptions specify larger, more complex Q-matrices with multiple rows for each item (Bradshaw & Templin, 2014; Kuo et al., 2018). This is likely to make accurate specification of the matrix more difficult. Models which allow for a smaller, less complex Q-matrix should also allow for better Q-matrix specification and more precise measures of misconceptions while keeping the test at a reasonable length for classroom use. This study proposed a test format that used just such a model and used it to develop a test.

Purpose and Research Questions

The purposes of this study were to demonstrate the construction of a cognitive diagnostic assessment to measure knowledge and misconceptions using a newly proposed test format, to investigate diagnostic cognitive model fit for responses to the instrument, and to use the instrument to investigate the relationship between ability and the presence of specific misconceptions. I did this by creating the Misconceptions About Force Assessment, an instrument that measures knowledge of and misconceptions about Newton's first and second laws. Knowledge was modeled as a unidimensional construct and estimated with an item response theory model and misconceptions were modeled as discrete skills and estimated with a diagnostic cognitive model. The structure of the instrument allowed identification of misconceptions even in students who answered physics knowledge questions correctly.

To address the purposes of this study, three research questions were examined:

1. How well do the specified psychometric models (item response theory and diagnostic cognitive model) fit responses to the instrument?
2. How do responses on this instrument compare to responses to Force Concept Inventory items which measure the same knowledge and misconceptions?
3. What is the relationship between physics knowledge and the presence of specific misconceptions as measured by this instrument?

Conceptual Framework

This study combined conceptual frameworks from educational psychology (e.g. constructivism), science education (e.g. conceptual change), and psychometrics (e.g. latent variable models). First, constructivism provided a way to think about the creation and structures of student knowledge. Second, conceptual change provided a framework within which to consider how student misconceptions develop and change. Third, the psychometrics field has developed multiple latent variable models and theories for measuring individual's knowledge and skills by interpreting their responses to assessments. The MAFA was designed using item response theory and a diagnostic cognitive model.

Constructivism is the prevalent theory of learning in education (Jones & Brader-Araje, 2002) and is used to refer to not only how students learn but the nature of knowledge itself. Constructivists believe that knowledge is constructed by individuals as they make sense of their experiences—both alone and through discussion with others (Foote et al., 2001). This has major implications for teaching. First, we cannot consider teaching to consist of the delivery of a way of thinking or a set of facts. Teaching and learning consist of meaning making. As they make meaning, students relate new information to what they already know. Because every student

interacts with lessons differently depending on their prior knowledge and experiences, different students will make different meaning of the same experiences. They will learn different things. Learning is not the simple transfer of a set of knowledge from teacher to learner. In fact, there is no set of knowledge independent of the individual. Constructivism helps to explain the prevalence of misconceptions in science. Although science teachers may agree on the knowledge that they are teaching, they cannot predict how their students will incorporate new ideas into their constructions nor what the final conception will be (Harrison & Treagust, 2001). Often, the final conception is very different from the expert knowledge that science teachers are trying to convey (Harrison & Treagust, 2001; Sadler, 1998; Schneps & Sadler, 1988; Vosniadu & Skopileti, 2014).

It is generally agreed that students enter the science classroom with a set of preexisting concepts and beliefs about how the world works (Brown & Hammer, 2013; Duit & Treagust, 2003; Vosniadou & Skopeliti, 2014). Many of these ideas are incorrect, resistant to change, and hinder students' progress in learning science (Halloun & Hestenes, 1985a; Hewson and Thorley, 1989). In a constructivist view of education, teachers structure experiences and discussion with the purpose of guiding student conceptions toward a closer fit with scientific thinking. "Conceptual change" is the process whereby learners restructure their existing concepts and beliefs to incorporate new knowledge (diSessa & Sherin, 1998; Duit & Treagust, 2003). Learning to think like a scientist involves adopting and understanding new sets of ideas (Posner et al., 1982), a process which occurs gradually and along paths that are not always readily visible. Rather than simply discarding their old ideas and adopting new ones, learners adapt their old ideas to incorporate new ones. The degree to which they are willing to adapt their old ideas is

highly variable (Harrison & Treagust, 2001; Vosniadou, 2014) and the intermediate models that they form often retain misconceptions stashed among scientific knowledge.

Finding out what students know about a topic usually requires interpreting their responses to assessments. Psychometrics provides theories and methods that are used to guide the process of using test responses to determine what students know and can do. A simple definition of measurement is the process of assigning "numerals to objects or events according to rules" (Stevens, 1951, p. 22), although this definition may be debated in the social sciences. The types of quantities that are measured in the physical sciences such as height and weight can be measured directly with a level of precision that depends mainly upon the measuring instrument. In psychology and education, one is generally interested in describing attributes that cannot be measured directly. These are called *latent variables* or *constructs*. Examples include math ability, Newtonian thinking, emotional stability, and business acumen. Such quantities can also be measured to some level of precision that depends upon the measuring instrument. The difference is that a case must be made that the instrument measures what it is intended to measure. While there is agreement that a person's height can be measured with a meter stick, there may be some disagreement about what a written test such as the SAT measures. This research involved designing an assessment which can be used to measure two latent variables—knowledge of Newtonian physics and the presence of misconceptions about Newtonian physics—and to provide feedback that can inform subsequent instructional decisions. Collecting evidence to support the use of the instrument for this purpose was an integral part of the test development process.

Overview of Methodology

The focus of this research was on the development of a diagnostic cognitive assessment. DiBello et al. (2007) list six components involved in the development of a diagnostic assessment:

1. Determining the purpose of the assessment;
2. Describing the latent skills that will be assessed;
3. Developing and analyzing the assessment items;
4. Specifying the psychometric model that represents the relationship between latent skills and item responses;
5. Estimating and evaluating the model and person parameters; and
6. Creating a system to report the results of the assessment to stakeholders.

They note that the process of test development is messy and that earlier steps will be revised as later steps are completed. In the following paragraphs I outline the actions I took to develop the MAFA and to answer the research questions.

The main purpose of the MAFA, to diagnose student misconceptions in Newtonian physics, was defined (step one) and its importance was supported by the literature described in earlier sections of this chapter. Step two, defining the latent skills (misconceptions in this case) to be assessed, was accomplished by researching the literature on students' physics misconceptions. According to DiBello et al. (2007), the decision of how to represent skills and attributes should be informed by the pertinent literature. Numerous studies have explored and described student misconceptions in Newtonian physics (Champagne et al., 1980; Clement, 1982; Gunstone & White, 1981; Halloun & Hestenes, 1985a, 1985b; McCloskey et al., 1983;

McCloskey et al., 1980; Minstrell, 1982; Trowbridge & McDermott, 1980, 1981; Viennot, 1979). This project used the prior research to define the physical situations and associated misconceptions that were included in the test domain. The authors of the FCI also referred to many of these earlier studies (Halloun & Hestenes, 1985a, 1985b). The focus of this study was not the identification of the universal set of misconceptions about Newtonian physics, but the individual diagnosis of a subset of misconceptions from that list. Step three, developing the assessment items, involved choosing physical situations to include in the assessment and crafting questions about them. Because given physical situations are linked to specific misconceptions, these were also primarily defined by previous research.

Two types of psychometric models were used (step four)—IRT models (the one-parameter and two-parameter logistic models) and a diagnostic cognitive model (DCM). The IRT model was used to measure students' mastery of Newtonian physics and the DCM was used to measure misconceptions. IRT models estimate individuals' placement along a continuous ability scale for some latent variable. For this assessment, the variable was mastery of Newtonian physics. In addition to measuring students' knowledge, a second purpose of the assessment was to classify students into classes based on the misconceptions they had about Newtonian physics. Cognitive diagnostic models, which are designed to measure which of a set of latent skills or attributes an individual has and has not mastered, are designed for this type of assessment. More details on each of these statistical models and the reasons for choosing each are provided in Chapter 2. Step five, the estimation of model parameters and fit statistics, is briefly described in the next paragraph under research question one. I did not develop a method for reporting test results to stakeholders (step 6) as part of this project. Although effective score reporting is a necessary part of developing any useful instrument, the focus of this research was on instrument

development, not score reporting. Creating such a method is, however, an important topic for future research.

Research question one involved determining the psychometric properties of the instrument. This was done in two phases. First, I conducted a pilot test with 100 university students. I analyzed the responses with classical test theory to determine how well the items performed and revised the items using the results of the analysis. Second, I conducted a field test with 349 respondents. I fit the item response theory and diagnostic cognitive models to the resulting data in order to estimate item parameters, person parameters, and overall model fit.

Research question two was part of gathering data for instrument validation, a process that is always involved in instrument development. One aspect of validity is construct validity, "the relationship between the content of a test and the construct it is intended to measure" (American Educational Research Association [AERA] et al., 2014, p. 14). I gathered data on construct validity by asking respondents to answer selected FCI questions which had distractors aligned to the misconceptions included in the MAFA. A comparison of responses to items which mapped onto the same misconceptions provided evidence for construct validity and to answer research question two. Finally, to answer research question three, I compared the item response theory based measure of physics knowledge to the frequency of each misconception.

Significance of Study

The results of this study have implications for educational measurement and for science education. In the field of educational measurement, the development of statistical models has outpaced their application to assessment development. For instance, psychometricians have developed multiple DCMs for measuring student misconceptions and some of these models have

been fitted to real data from conventional tests. However, few CDAs have been developed specifically for DCMs and there is little information in the literature on methods for doing so. This study proposed a test structure and demonstrated a method for developing CDAs to simultaneously measure student knowledge and misconceptions. It identified some of the difficulties inherent to the development of future CDAs (e.g. developing guidelines for which items to eliminate and which to keep) which are important areas for future research. The results of the research can inform the construction of future DCM-based concept inventories which are psychometrically sound.

In the field of science education, concept inventories are widely used to measure student knowledge and to judge the effectiveness of teaching. Most concept inventories are based on CTT and measure student knowledge as a single score. Concept inventories which provide a single score do not provide enough information about student knowledge to instructors. Metanalyses show that two of the most influential influences on academic achievement are conceptual change programs and the use of formative evaluation (Hattie, 2015). More fine-grained information about student knowledge, such as whether a student possesses specific misconceptions, can be used formatively to directly address misconceptions through instruction. This study provided a method of using existing research on student misconceptions and a test format that can be used to develop tools that provide more fine-grained information about student knowledge and misconceptions. The application of the method to develop new concept inventories that follow the new format is an area for future research.

Finally, the study has implications for physics education. US physics students underperform students from other countries. Multiple factors contribute to this situation. Two factors are: 1) the severe, long-term shortage of qualified physics teachers in the US, and 2) the

retention of misconceptions by many students even after completing a course in physics. Many physics teachers (lesser qualified teachers in particular) may not have the background to recognize student misconceptions. Even teachers who have the ability may not have the time to identify and address individual misconceptions in the classroom. The MAFA, a psychometrically sound tool that can be used by all physics teachers and students to identify misconceptions, was developed as part of this project. In addition to classroom use, the MAFA can serve as a tool for education researchers to test the effect of interventions on conceptual change.

Delimitations

Chapter 1 has served to place this study in multiple contexts: 1) the current and historical contexts of STEM education in the US, 2) the context of current research in educational measurement, and 3) the context of conceptual change research. The chapter began with a description of the current call to improve the quality of STEM education to prepare all citizens to participate in an innovation economy and to ensure that the US remains competitive within the global economy. It showed the significance of physics education to meeting this call. Next, the chapter described the gap between the development of psychometric models for measuring misconceptions and their application to the development of cognitive diagnostic assessments. Finally, it provided a brief description of the research that was conducted as well as the conceptual frameworks that informed the research. Next, I describe the delimitations of the study and why they were chosen.

This study was bounded by the decisions that were made in the design phase of the project. First, it did not attempt to identify student's misconceptions about Newton's laws directly. Instead, it used previous research about student misconceptions to develop test items. The focus of the study was to develop a cognitive diagnostic assessment and I decided to use the

rich collection of existing research about student's misconceptions to build the diagnostic tool. Second, I chose a limited number of common misconceptions to include in the assessment. Although there are many additional misconceptions that students might have about Newton's laws, including them in the assessment would have required the development of a longer test and/or a larger sample size. A longer test would have been less practical as a tool for formative assessment and a larger sample size would have been difficult to accomplish. Third, I chose to include only the first two of Newton's three laws of motion in the test domain. Including the third law would have meant replacing some of the misconceptions about the first and second laws with misconceptions about the third law. Of the three laws, I found that misconceptions about Newton's third law were the easiest to identify and address as a physics teacher. Therefore, I decided to exclude knowledge and misconceptions about the third law from the test domain. Fourth, I chose to limit the participant pool to undergraduate students who had completed no more than two semesters of college level physics. This choice was made because I hoped to investigate the prevalence of misconceptions in students who had varied levels of physics knowledge. I chose not to include students who had taken more physics courses because I believed that the likelihood of possessing misconceptions would continue to decrease with further physics instruction. Fifth, I fit only one DCM to the response data. There may be different DCMs that would show better fit, but because this was a proof-of-concept study to demonstrate how a DCM could be used, I decided to test only one model. Other methodological limitations included the representativeness of the sample (students self-selected to participate), the length of the test (just long enough to fit the 2-PL model given the sample size), and the software packages that were used to estimate the statistical models—different software packages produced different values for the standard error.

The remaining chapters of the paper are as follows. Chapter 2 provides a literature review of salient topics including research on student misconceptions in science, conceptual change, concept inventories, item response theory, and diagnostic cognitive models. Chapter 3 is the first of two manuscripts and addresses the test development process. Chapter 4 consists of the second manuscript which uses MAFA response data to investigate the relationship between knowledge level, physics education, and the possession of misconceptions about Newton's laws. Chapter 5, the final chapter, summarizes the research findings, places them within the larger context of the literature review and describes directions for further research and limitations of the study.

Chapter 2

Literature Review

In this study I developed an assessment to measure student knowledge and misconceptions that is scored using current psychometric models—a *cognitive diagnostic assessment* (*CDA*). Developing such an assessment requires knowledge of the topic being assessed, how the topic can be presented, and how test responses are used to judge student knowledge. Therefore, literature from both science education and psychometrics was necessary to provide a background. There are many attempts to measure the prevalence of student misconceptions in science education, but few tests provide profiles of students' misconceptions. At the same time, psychometricians have developed multiple statistical models which could be used to provide these profiles. These models have been tested with simulated data and used to model a few existing tests, but new assessments have not been developed to take advantage of them. Pertinent research in science education addresses student misconceptions, how these develop and change, and how they are measured while pertinent information from psychometrics addresses models that can be used to gauge student knowledge and the possession of misconceptions. In the first part of this chapter, I present information about misconceptions in science, how misconceptions relate to conceptual change and student learning, and how knowledge of them might inform instruction. Next, I present information about concept inventories with an emphasis on how researchers have analyzed responses from commonly used concept inventories in physics. In the final section of this chapter, I present information about the two types of measurement models that are used in this study—item response theory (IRT) and diagnostic cognitive models (DCMs). The information will help to provide a background in

measurement for readers who are experts in science education and a background in science education for those who specialize in educational measurement.

Science Education

Students do not always know everything their teachers believe they do. Two poignant illustrations of the disconnect that may exist between what teachers think they have taught their students and what their students actually know are found in the film "A Private Universe" (Schneps & Sadler, 1988). In the first, an interviewer asks MIT engineering students to light a small bulb with a battery and one wire. He finds that few students can accomplish this simple task—one that most engineering professors would probably assume their students could perform. In the second, a ninth-grade Earth Science student—one of the brightest according to her teacher--is asked to explain the seasons. At first, her explanation of indirect sunlight appears to agree with the scientific explanation that the changing angle of the sun causes temperature changes. When asked to draw a picture to show what she means by indirect sunlight, however, it becomes clear that the student has an alternate conception. She draws rays of sunlight bouncing off an indeterminant point in space to reach Earth. In both of these cases, teachers are surprised to find that students who have succeeded and even excelled on class assessments have failed to understand what was taught. This disparity has been confirmed by other researchers in science education (Diakidoy & Iordanou, 2003; Mazur, 2009) and reading and math education (Eckert et al., 2006).

Researchers who study students' ideas about science use many different terms to refer to them. Some examples are: phenomenological primitives (p-prims) (diSessa, 1993), preconceptions (Clement, 1982; Halloun & Hestenes, 1985b), alternative conceptions (Chi et al.,

1981; Viennot, 1985), framework theories (Vosniadou, 2014), and misconceptions (Halloun & Hestenes, 1985a; McCloskey et al., 1983). Some of these terms are used to differentiate between correct and incorrect ideas. For instance, *preconceptions*, ideas that exist before instruction, may be correct or incorrect. Some researchers use the term *alternative conception*, which may be a mix of correct and incorrect scientific ideas, to recognize the usefulness of the conceptions to students and the possibility that teachers may leverage alternative conceptions to increase the effectiveness of instruction. Other terms are used to express nuanced meanings about the organizational complexity of student ideas and how they compare to scientific concepts. For instance, Vosniadou and Skopeliti (2014) consider naïve physics to be a "framework theory"—a loosely structured set of related concepts that are based on everyday culture and experience and are rooted in a set of ontological beliefs. These are usually private and held to a lower standard of forecasting power and internal consistency than are scientific theories. In contrast, diSessa's (1993) p-prims are small pieces of knowledge or beliefs that develop from everyday experience and are activated in the interpretation of new experiences. These nuanced definitions may be useful when explaining the mechanisms whereby conceptual change occurs. For the purpose of designing assessments, however, a different classification system may be more useful.

In *Science Teaching Reconsidered: A Handbook* (NRC, 1997), the authors classify students' nonscientific ideas about science into five categories:

1. Preconceived notions—These are conceptions that have been developed from common experience. For instance, students may believe that more massive objects exert greater forces than less massive objects.
2. Nonscientific beliefs—These are ideas, such as belief in a young Earth or intelligent design, that have been learned from nonscientific sources.

3. Conceptual misunderstandings—These are schemas that students have constructed by melding pieces of scientific knowledge into existing beliefs.
4. Vernacular misconceptions—These are words that have a different meaning when used scientifically than in everyday situations. Two examples are the use of "work" in physics and the use of "indirect" when referring to the angle of sunlight in earth science.
5. Factual misconceptions—These are falsehoods that were learned (sometimes from a poorly informed teacher) and never corrected such as the idea that wave motion causes particles to have a net displacement.

Some types of misconceptions may be more easily changed than others—a difference that may be important when designing instruction. Misconceptions that are incorporated into students' personal conceptions about the world may be more difficult to dispel than vernacular or factual misconceptions. Conceptual change involves changing beliefs and ideas (diSessa & Sherin, 1998). These involve deep ways of thinking that go far beyond memorization of facts.

Many teachers assess what students already know prior to instruction to determine what parts of this knowledge are scientifically correct. I will adopt Lucariello and Naff's (n.d.) terminology to refer to this prior knowledge. I will refer to all pre-instructional knowledge as *preconceptions*. I will call preconceptions that agree with scientific theories and facts *anchoring conceptions* and those that do not agree *misconceptions*. It is important for teachers to understand both types of students' preconceptions because instruction should vary depending on whether they contradict or agree with established scientific ideas. The application of theories of *constructivism* to education has resulted in the generally accepted model of learning in which

students *construct* knowledge by connecting new knowledge and experiences to their preconceptions (Jones & Brader-Araje, 2002; Lucariello & Naff, n. d.; Matthews, 1998; NRC, 1997; Phillips, 1995). Because they do not simply replace their existing ideas with new knowledge, delivering scientifically correct information will not always result in students developing scientifically accurate models of the world. To develop scientific conceptions, students must confront their misconceptions and reconstruct their mental models (NRC, 1997). When they incorporate new knowledge with misconceptions, students may combine the two for a partially correct model or they may distort the scientific information for a completely incorrect model. In either case, it may be useful for teachers to gauge student preconceptions as they plan instruction.

Constructivism in Education

The wide and significant influence of constructivist theories on education may have been a backlash against the earlier, widespread adoption of behaviorist practices by school administrators (Jones & Brader-Araje, 2002). In the 1960s, the application of behaviorism to teaching resulted in a system in which teachers delivered information to students who then learned the information through practice. There was little emphasis on probing how students developed new knowledge. It was believed that if teachers provided the correct learning activities and then used positive and negative reinforcement effectively, students would learn. This practice failed to deliver the desired results and, in the 1970s, constructivism began to become more widespread as a theory to explain how students learn (Jones & Brader-Araje, 2002).

There are many versions of constructivism. One of the most influential forms in education is Vygotsky's social constructivism (Jones & Brader-Araje, 2002). Social

constructivism posits that student discourse is an essential component of knowledge construction. Its influence is widespread and seen in teaching strategies that employ carefully guided group work, discussion, and debate—practices that encourage students to make meaning with each other. Matthews (1998) distinguishes constructivism in education, which focuses on how knowledge is formed, from philosophical constructivism in science, which focuses on what counts as scientific knowledge. In its most applied form, educational constructivism may concern only the nature of pedagogy that promotes the construction of sound scientific knowledge. Phillips (1995) provides a simple definition of constructivism when he states:

> Undoubtedly, humans are born with *some* cognitive or epistemological equipment or potentialities…, but by and large, human knowledge, and the criteria we use in our inquiries are all *constructed* (p.5).

For the purposes of this research, we might consider a definition of constructivism that lies somewhere between Phillips and pedagogical constructivism. That is, we may consider that the process of student learning is one in which students construct new knowledge by linking and adapting what they learn to what they already know and that constructive pedagogy considers students' existing knowledge during instructional decision making.

The move toward a constructivist view of learning in education coincided with an increased interest in student preconceptions. There was a rapid increase in the number of publications concerning students' pre-instructional perceptions in science from the 1970s to the 1990s with the heaviest interest on students' preconceptions about physics topics (Duit, 1993). Researchers gathered data using many methods such as informational interviews (Clement, 1982; Novick & Nussbaum, 1978), both open-ended and constrained response written assessments

(Champagne et al., 1980), and observations of students working with lab equipment to perform an assigned task (McDermott, 1984). The classifications of common misconceptions from these studies have been used to develop *concept inventories*. Concept inventories are typically multiple-choice assessments for which the distractors correspond to common misconceptions about the topic.

Concept Inventories

For science topics in which common misconceptions have been identified, concept inventories provide a much more efficient way to collect evidence of student preconceptions than student interviews or analysis of student work. Many examples of concept inventories exist. Some examples are listed in Table 2.1. Almost all of the concept inventories that I found were developed using CTT, although a few have been scored retroactively using different methods. Although the distractors correspond to common misconceptions, inventories are not typically scored to take advantage of this. Concept inventories are often used as pre- and post-tests with the change in scores calculated as normalized gain and the mean class values used to compare performance under different conditions (Hake, 1998; LoPresto & Murrell, 2011; Thornton et al., 2009; Williamson et al., 2016; Yeo & Zadnick, 2001). Distractors are written to be appealing to students who have misconceptions. When scores increase after instruction, it may be interpreted to mean that students have fewer misconceptions. However, this type of scoring does not provide information about the frequency of specific misconceptions which makes it difficult to measure how they may have changed during instruction.

Retrofitting Concept Inventories with New Psychometric Models

Alternate scoring methods have been applied to some of these assessments after they

Table 2.1

Examples of Concept Inventories and Scoring Methods by Discipline

Discipline	Name and Reference	Scoring
Physics	Force Concept Inventory (Hestenes et al., 1992)	CTT
	Mechanics Baseline Inventory (Hestenes & Wells, 1992)	CTT
	Force and Motion Conceptual Evaluation (Thornton & Sokoloff, 1998)	CTT
	Brief Electricity and Magnetism Assessment (Ding et al., 2006)	CTT
	Thermal Concept Evaluation (Yeo & Zadnick, 2001)	CTT
	Newtonian Gravity Concept Inventory (Williamson, 2013)	CTT & IRT
Chemistry	Chemistry Concept Inventory (Pavelich, et al., 2004)	CTT
	Quantum Chemistry Concept Inventory (Dick-Perez et al., 2016)	CTT
Astronomy	Light and Spectroscopy Concept Inventory (Bardar et al., 2006)	CTT
	Astronomy and Space Science Concept Inventory (Sadler et al., 2010)	CTT
	Astronomical Misconceptions Survey (LoPresto & Murrell, 2011)	CTT
	Conceptual Inventory of Natural Selection (Anderson et al.,	
Biology	2002)	CTT
	Biology Concept Inventory (Klymkowsky et al., 2003)	CTT
Geoscience	Climate Change Inventory (Jarrett et al., 2012)	CTT
	Geoscience Concept Inventory (Libarkin & Anderson, 2005)	IRT

were developed. For instance, IRT models have been fitted to the FCI (Planinic et al., 2010; Wang & Bao, 2010) and the FMCE (Talbot, 2013). The use of IRT allows for more readily comparable scores between different groups of respondents, but it does not provide direct information about student misconceptions. At least two groups of physics education researchers have developed methods to analyze incorrect responses for evidence of misconceptions and applied these to the FCI and FMCE. Next, I describe two of these efforts for the FCI which is the most widely used concept inventory in physics education (Smith & Tanner, 2010).

In the first example, researchers analyzed incorrect responses on the FCI to compare the prevalence of certain misconceptions between groups of first-year engineering students (Martin-

Blas et al., 2010). First, the researchers determined if there was a dominant incorrect response to each question. This was defined as a response that accounted for at least half of the total incorrect responses. Second, each dominant incorrect response was aligned with a misconception as defined by the FCI authors (Hestenes et al., 1992). Then the frequency of the dominant misconceptions was compared between two groups of students. In the second example, Fulmer (2015) treated FCI questions as ordered multiple-choice items by mapping incorrect responses to positions on two learning progressions of force and motion. Lower levels on the learning progressions represented a greater prevalence of misconceptions related to a topic. The researcher then modeled the data using a rating scale Rasch model. There was only moderate model fit and it was difficult to distinguish between performance levels. Although the researcher found high interrater agreement between experts who mapped the items onto the learning progressions, it may be that the misconceptions on the FCI do not fit the learning progressions well. It should also be noted that some developers of DCMs have used sets of FCI response data to test their models (Bradshaw & Templin, 2014; Kuo et al., 2018). I provide more detail about these studies in a later section of this chapter.

Modeling Test Data

What is Being Measured

Educational measurement involves using test data to make inferences and decisions about individuals or groups by linking test responses to a scale using a statistical model. Test content is designed to measure the desired quantity. This varies depending on the inferences or decisions that will be made. For instance, tests have long been used to measure content mastery separately from cognitive processes (Osterlind, 2010). Today, educators tend to measure quantities such as

reading or math ability rather than simple content knowledge. The characteristics of people that need to be measured in education are referred to as *latent variables* because, unlike physical quantities such as length and volume, they cannot be measured directly. Learning to read or to do math is complex and involves multiple cognitive processes such as interacting with stimuli, processing information from the interaction, and making sense of the information by constructing a mental model and mapping it onto an existing network of knowledge (Osterlind, 2010). When we measure knowledge and skills, we theorize that students will call upon these same cognitive processes as they respond to test items and that, by interpreting their responses, we can quantify the extent to which the student is performing the process.

The terms *latent variable* and *construct* are often used interchangeably in the literature on testing. By definition, a latent variable is unobservable. It is a cognitive state or process which exists or occurs within a person's brain. A construct can be considered a "meaningful description of a particular psychological trait or latency" (Osterlind, 2010, p. 4). Hence, a construct is an attempt to make that which cannot be observed manifest. In educational measurement, we are often concerned with measuring students' knowledge of a given subject. I will use the term *ability* to refer to this quantity as is a common habit in educational measurement. The term does not imply that student knowledge is inherent and unchangeable as it might in common use. In fact, ability by this definition is the thing that instruction is designed to change. Usually, ability is measured on a continuous scale. Other terms that are seen in the literature are *skills* and *attributes*. These are often used to refer to student knowledge that is measured on an ordinal scale and at a finer grain size than ability. For instance, one's math ability may depend on multiple discrete skills such as adding whole numbers and multiplying decimals.

Approaches to Testing

Three approaches to interpreting test results are classical test theory (CTT), item response theory (IRT), and latent class analysis (LCA). Each approach has its own set of assumptions that should be met and parameters which can be estimated. One large-scale difference is that CTT and IRT model ability as a continuous variable, while latent variable analysis models it as categorical variables. A second difference is that in CTT and most uses of IRT it is assumed that all test items measure the same latent variable and individuals' responses are used to estimate their ability in that variable along some continuous scale. In contrast, using LCA to analyze test responses assumes that the test items measure multiple constructs and individuals' responses are used to place them in a group (a "class") with individuals who have a similar pattern of constructs. A third difference is that CTT uses total test scores to make inferences about a construct and IRT and LCA analyze responses at the item level. This research focuses primarily on IRT and cognitive diagnostic models (a type of latent class analysis) which will be presented in following sections of the paper. CTT, which will be used to test item quality using data from the pilot test, will be compared to IRT.

Item Response Theory

Background

Despite its name, item response theory (IRT) is not a theory in the scientific sense. It is a group of probabilistic latent variable models—statistical models that relate probabilistic measures of latent variables to sets of observations designed to measure them. For our purposes, the latent variables will be measures of ability (application of Newton's laws) and the observations will be responses to test items. There are IRT models for polytomous items, but

because I used dichotomous items only, I limit this discussion to those for dichotomous items. These models can be separated into two categories, the Rasch model and item response curve models. Although the Rasch model is theoretically different from the simplest item response curve model, they differ mathematically only by a scaling constant. Because this project did concern the theoretical differences, this discussion is limited to item response curve models.

Latent variable models were introduced by Frederic Lord in the early 1950s but were not widely used in educational measurement and test development until much later (Hambleton & Cook, 1977). Application of the models to test development was slow, but by the late 1970s, test developers were beginning to use the models to design tests and to explain student responses (Hambleton & Cook, 1977). Reasons that latent variable models were not adapted more quickly include their mathematical complexity, lack of convenient estimation software, the difficulty of satisfying the model assumptions, and the uncertainty of the robustness of the model estimates to violations of assumptions (Hambleton & Cook, 1977; Osterlind, 2010). The popularity of the models has grown as software has become more readily available and as IRT-focused literature has expanded beyond its initial theoretical focus. Today, there are multiple software packages that can be used to estimate various IRT models and books on the subject that are approachable by non-psychometricians. Many large-scale tests are scaled using IRT.

IRT and CTT Compared

In classical test theory, a person's observed score on a test is assumed to be the sum of their true score and an error term. The observed score is generally the sum of the points for correct responses while the true score-- the score that a person would have if they were to answer every possible item about the topic an infinite number of times—cannot be measured. Because the error term (the difference between the unknown true score and the actual score on the test) is

random, it will get closer to zero as the test gets longer. Therefore, longer tests will generally produce more accurate scores. In contrast, it is possible for IRT-based tests to provide an unbiased estimate of examinee ability with very few items. This is a distinct advantage over CTT-based tests for many uses (Hambleton & Cook, 1977). One reason that CTT continues to be used is that test statistics can be estimated with much smaller sample sizes (fewer respondents) than for IRT. This is the reason that I evaluated the quality of items on the MAFA using CTT after the pilot test. I did not have enough responses to use IRT at this point.

There are a number of issues with using CTT that can be minimized or eliminated by using IRT. One of the greatest limitations of CTT is that item difficulty and item discrimination depend upon who takes the test. In CTT item difficulty is defined as the percent of respondents who answer an item correctly. The same item may have a low difficulty with students of high ability and a high difficulty with students of low ability. This is at odds with the idea of latent variables which "occupy a latent space that can be quantified along a hypothesized infinity continuum from $(-\infty, \infty)$" (Osterlind, 2010, p. 273). Item response models solve this problem by measuring item difficulty and person ability on the same scale. In the 1-PL and 2-PL IRT models, the difficulty of a dichotomous item is defined as the ability at which an individual has a 50% chance of answering the item correctly. This is halfway between the probability of a correct response for a person with ability of $-\infty$--a probability of 0.0--and for a person with an ability of $+\infty$--a probability of 1.0. In the 3-PL IRT model item difficulty and ability are also measured on the same scale, but the probability of a person with ability of $-\infty$ having a correct response is greater than zero. Therefore, the difficulty for the 3-PL model is defined as the ability for which the probability of answering correctly is halfway between the probability of guessing correctly and 1.0. A second problem is that CTT models the standard error of measurement (SEM) as

constant across all levels of ability. (The *standard error of measurement* is the standard deviation of the measurement error on the test.) This may not be accurate—especially for very low or very high abilities. In contrast, item response theory allows the SEM to vary with ability level. Finally, in CTT reliability is described at the test level and the assumptions required to calculate a measure of reliability (some version of parallel test forms) may be difficult to meet. Item response theory approaches reliability with the test information function—a measure of the abilities for which the test gives the most precise estimates.

The Main IRT Models

The three most commonly-used IRT models are the one-parameter logistic (1-PL), two-parameter logistic (2-PL), and three-parameter logistic (3-PL) models (Osterlind, 2010). Each of the three main logistic IRT models for dichotomous responses provides an estimate of person ability along a random scale that is also used to measure item difficulty. It is assumed that the ability being measured is approximately normally distributed within the population and IRT software generally scales ability along the standard normal distribution, N(0,1). The names of the three models reflect how many item parameters are allowed to vary. In the one-parameter logistic (1-PL) model, items are allowed to differ only on the *difficulty* parameter. In IRT, *item difficulty* is the value at which a person of that ability has a 0.5 probability of answering the item correctly when the probability of guessing the correct answer is modeled as zero as it is in the 1- and 2-PL models. For instance, if an item has a difficulty of 1.3 according to these models, then the probability of a person with ability of 1.3 answering correctly is 0.5. The probability is lower for a person with lower abilities and higher for persons with higher abilities. The two-parameter logistic (2-PL) model also allows the *item discrimination* parameter to vary. Item discrimination is a measure of how quickly the probability of a correct response changes as ability changes.

Higher discriminations correspond to a higher rate of change. The three-parameter logistic (3-PL) model adds a *guessing parameter* (the probability of a respondent with ability of - ∞ answering correctly) for each item. In the 1-PL and 2-PL models the guessing parameter is taken to be 0.

Assumptions of Item Response Theory

Because the mathematical development of IRT models is based on assumptions about the data, it is important to consider whether the assumptions have been met. Osterlind (2010) lists three important assumptions that should be met. First is the assumption of "unidimensionality". In a unidimensional model, all test items measure the same, single latent variable. This can be tested using factor analysis. The subject of this research, the Misconceptions About Force Assessment (MAFA), measures knowledge of Newton's first and second laws. A second assumption is that items are locally independent. Responses to all items should depend only on the latent variable being measured. Unusually high correlations between item responses after controlling for person abilities (i.e. correlations between residuals) may indicate that this is not the case. Such correlations may occur for items that share information such as a reading passage or diagram. A third, important assumption is that the model fits the data. Important factors to consider when choosing an appropriate model include test length and sample size. As model parameters are added, some combination of a larger sample size and/or a longer test is required for estimation. Despite the abundance of research concerning the effects of these factors on model fit and parameter estimation, there are no exact guidelines. However, recommendations suggest that for the 2-PL model, a 10-item test should have a minimum sample size of 750 (Alpir & Duygu, 2017) while a 20-item test should have a sample size of about 500 (Alpir & Duygu, 2017; de Ayala, 2009). Yen and Fitzpatrick (2006) also note that shorter tests and smaller sample

sizes may be used in low stakes applications such as during field testing. Recommendations are lower for the 1-PL model and higher for the 3-PL model. Because my sample sizes were relatively small for IRT, I tested the 1-PL and the 2-PL models which have fewer parameters than the 3-PL model. Osterlind (2010) refers to a fourth assumption, which applies to all types of assessment. This is the idea that respondents apply their ability to every item. There is no part of the statistical model that accounts for respondents using less than their maximum ability on any item and certainly no easy way to test that they have. In the next section, I describe the 1-PL and 2-PL models.

The 1-PL and 2-PL Models

The 1-PL model relates the probability of a correct response on item *j* to person ability (θ), item difficulty (β), and item discrimination (α) as:

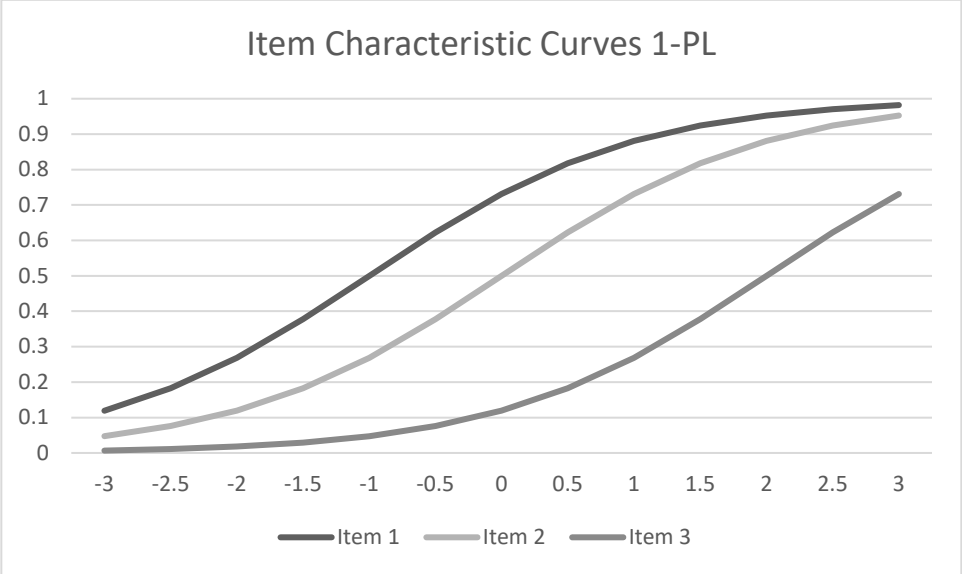
$$p(x_j = 1|\theta, \alpha, \beta_j) = \frac{e^{\alpha(\theta-\beta_j)}}{1 + e^{\alpha(\theta-\beta_j)}} \quad (1)$$

where $x_j = 1$ indicates a correct response to item *j* and $x_j = 0$ indicates an incorrect response. When shown graphically, this function (called the *item response function*) produces a characteristic plot called the *item characteristic curve* (ICC). It should be noted that the item discrimination (α) is constant for all items in this model. Figure 2.1 illustrates the item characteristic curves for three items with an item discrimination of 1 ($\alpha_1 = \alpha_2 = \alpha_3 = 1$) and with item difficulties of -1, 0, and 2, respectively ($\beta_1 = -1$, $\beta_2 = 0$, and $\beta_3 = 2$). It can be seen that the probability of a correct response is 0.50 when person ability equals item difficulty and that this probability decreases for lower abilities and increases for higher abilities. It is a monotonically

increasing function. The curve is asymptotic such that the probability approaches zero as person ability approaches -∞ and one as person ability approaches +∞. The item discrimination parameter is related to the slope of the curve at the inflection point with larger discriminations producing steeper slopes. Because the discrimination is the same for all items in this model, the ICCs do not intersect.

Figure 2.1

ICCs for 1-PL Model



Note. The x-axis measures ability and item difficulty. The y-axis is the probability of a correct answer. The probability of a correct response is 0.5 when ability equals item difficulty. IRCs do not cross because they have equal discriminations.

The Birnbaum 2-PL model (Birnbaum, 1968) differs from the 1-PL model by allowing the item discrimination to vary between items so that α becomes α_j. This produces the equation

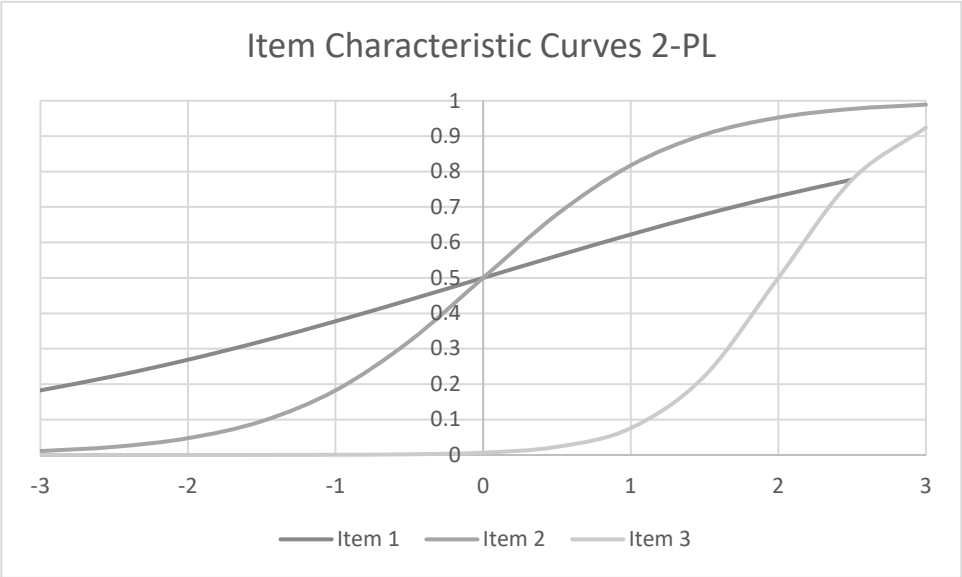
$$p(x_j = 1|\theta, \alpha_j, \beta_j) = \frac{e^{\alpha_j(\theta-\beta_j)}}{1 + e^{\alpha_j(\theta-\beta_j)}} \quad (2)$$

where α_j is the discrimination for item *j* and all other variables are defined as in the 1-PL model. The effect of varying item discriminations on ICCs is illustrated in Figure 2.2. Item difficulties are equal to 0 for the first two items and 2 for the third item ($\beta_1 = \beta_2 = 0$ and $\beta_3 = 2$). As before, varying item difficulties shifts the curves to the right or left. The item discriminations are allowed to vary in the 2-PL model. For this example, each item has a different discrimination ($\alpha_1 = 0.5$, $\alpha_2 = 1.5$, and $\alpha_3 = 2.5$). Greater item discriminations cause steeper maximum slopes—a greater increase in the probability of a correct response for the same increase in ability.

Both the 1-PL and 2-PL models represent person ability as the ability for which there is the greatest probability of producing the overall response pattern. Because item responses are locally independent, the probability of a set of responses is the product of the probability of the response to each individual item. If we write the probability of a correct response to a dichotomous item *j* as p_j, then the probability of an incorrect response can be written as $(1 - p_j)$. Thus the probability of correct responses to the first three items and an incorrect response to the fourth item on a four item test is written $p(x_1 = 1) * p(x_2 = 1) * p(x_3 = 1) * p(x_4 = 0) = p_1 * p_2 * p_3 * (1 - p_4)$. If item difficulties and discriminations are known, then it is a simple matter to use Equation 1 or 2 to calculate p_j for a given ability. In practice, however, neither person ability nor item parameters are known and they must be estimated together in a stepwise fashion. The general equation for estimating the ability of person *i* from a set of *J* responses is

Figure 2.2

ICCs for 2-PL Model



Note. The x-axis measures ability and item difficulty. The y-axis is the probability of a correct answer. The probability of a correct response is 0.5 when ability equals item difficulty. Varying item discriminations ($\alpha_1 = 0.5$, $\alpha_2 = 1.5$, $\alpha_3 = 2.5$) correspond to varying slopes.

written as the likelihood function:

$$L(\boldsymbol{x_i} \mid \theta, \alpha, \boldsymbol{\beta}) = \prod_{j=1}^{J} p_j^{x_{ij}} (1 - p_j)^{(1 - x_{ij})} \tag{3}$$

where $\boldsymbol{x_i}$ is the vector of person *i*'s responses, θ is person ability, α is item discrimination (constant for the 1-PL model and a vector of values for the 2-PL model), $\boldsymbol{\beta}$ is a vector of item difficulties, p_j is the probability of a correct response to item j for a person of ability θ, and x_{ij} is person *i*'s

response to item j. As more items are added, the value of the likelihood function becomes smaller. To prevent working with very small numbers, the practice is to take the natural log of the likelihood function, called the log likelihood function which is given by

$$\ln(L(\boldsymbol{x_i}|\theta,\alpha,\boldsymbol{\beta})) = \sum_{j=1}^{J}(x^{ij}\ln(p_j) + (1-x_{ij})\ln(1-p_j) \quad (4)$$

where p_j is calculated using either Equation 1 or 2. If item parameters are known, then Equation 4 can be used to calculate the value of the log likelihood function for each value of θ. The ability for which the log likelihood function takes the greatest value is the ability estimate for the response pattern $\boldsymbol{x_i}$. As mentioned earlier, neither item parameters nor person ability are generally known. They can be estimated together in a stepwise fashion using joint maximum likelihood estimation (JMLE). However, JMLE cannot provide ability estimates for either perfect or zero scores and JMLE estimates may be biased, especially for shorter tests. A solution to these problems is to use marginal maximum likelihood estimation (MMLE) in which item parameters are estimated using a theoretical ability distribution. In the next section, I explain the methods that were used in this study.

IRT Parameter Estimation

Item parameters and person ability estimates were estimated using two different methods. Item parameters were estimated with marginal maximum likelihood estimation (MMLE) using the expectation-maximization (EM) algorithm (Dempster et al., 1977) and person abilities were estimated using Bayes estimation, also called expected a posteriori (EAP) estimation. As stated earlier, estimation of one set of parameters (either item or person) requires knowledge of the

other set of parameters. Both methods solve this problem by starting with an estimate of the parameters that gets refined through an iterative process.

To calculate the unconditional probabilities of particular response patterns from a random sample of the population, it is necessary to integrate the probability function. Bock and Aitkin (1981) approximate the integral using Gauss-Hermite quadrature. It is assumed that the distribution of abilities within the population follows some known distribution (such as the standard normal distribution). The sample of person abilities that comprise the test responses are not assumed to follow this distribution. They simply form a finite number of points which might be found anywhere along the distribution. It is possible that they will all be in the low end or at the high end and MMLE estimates are robust to this condition. Instead of using these empirically situated points to estimate item parameters along a continuous distribution, a number of points (quadrature points) are chosen along the length of the abscissa. In this application, the quadrature points are ability values. Each quadrature point has an associated weight and the integral is approximated as the sum of the product of ordinate values and their weights. In this way, an approximation of the probability of each set of item responses is calculated for each quadrature point. This forms the "E" part of the EM algorithm.

Next, the ability estimates from the first step are used to estimate item parameters from the log-likelihood function. This is the "M" part of the EM algorithm. These item parameters are then used to estimate person abilities from the likelihood function. At each iteration, of the "M" step, the current item parameter estimates are compared to the prior estimates. The model reaches *convergence* when the difference between estimates becomes smaller than a specified threshold value. At this point the process ends with a set of item parameter estimates.

Using the EAP method to estimate person abilities for each response pattern does not require iteration but it does begin by establishing a hypothetical distribution (such as the standard normal distribution) of ability estimates for the population. If we define the density of the distribution curve as $F(\theta)$, then the weight for a given ability is given by

$$W(\theta) = \frac{F(\theta)}{\sum F(\theta)} \tag{5}$$

such that the weights sum to 1. In EAP the ability estimate is defined as the weighted mean of the posterior distribution of θ given the distribution of response patterns, x_i. The posterior distribution is given by the likelihood function, $L(\theta)$. This can be expressed as

$$\hat{\theta} = \frac{\sum \theta * L(\theta) * W(\theta)}{\sum L(\theta) * W(\theta)} \tag{6}$$

where $\hat{\theta}$ is the ability estimate. As in MMLE, a number of quadrature points are chosen along the hypothetical distribution of abilities and the summations are performed across these points. The EAP estimate exists for all response patterns and has the smallest standard error of any estimation method. Although EAP estimates tend to be biased toward the mean of the hypothetical distribution, this effect is generally small within $\pm 3\sigma$ of the mean as long as the posterior standard deviation is small (Cai et al., 2017).

IRT Model Fit

In IRT, model fit should be considered for both individual items and the overall model. For individual items, unusually large standard errors (greater than 1.0) may indicate problems. Because the 1-PL and 2-PL models are nested models, their fit can be compared using the log-likelihood function, *L*, as well as the Akaike Information Criterion (AIC) (Akaike, 1974) and the

Bayesian Information Criterion (BIC) (Schwartz, 1978). The Likelihood ratio test compares the value of $-2\ln(L)$ which follows the chi-square distribution with degrees of freedom equal to the change in number of parameters between models. However, the chi-square test is sensitive to large sample sizes which are needed to estimate IRT models. Another problem is that the value of the log-likelihood function generally decreases as more parameters are added to the model and this can lead toward overfitting. Both the AIC and BIC account for this by adding a factor for number of parameters. The BIC weights the number of parameters by the natural log of the sample size. The AIC and BIC are given by

$$AIC = -2\ln(L) + 2(k) \tag{7}$$

$$BIC = -2\ln(L) + \ln(n) * k \tag{8}$$

where *k* is the number of parameters estimated by the model and *n* is the sample size. A smaller value for either criterion indicates better model fit so these measures of model fit effectively penalize adding parameters to better fit the data.

Diagnostic Cognitive Models

Introduction

Diagnostic assessments in education may provide information about student's attributes which teachers can use to individualize instruction. Diagnostic cognitive models (DCMs) provide one way to interpret responses to diagnostic assessments. DCMs are confirmatory multidimensional restricted latent-class models in which each latent class is a profile showing whether a person possesses each of a set of *attributes*. Attributes are latent states that are needed

to respond correctly to the items on the assessment. Single items may require one or more attributes to answer correctly. The number of latent classes is restricted by the number of attributes. These are specified a priori. If the number of attributes specified in the assessment is A, then the number of possible latent classes (*N*) is given by $N = 2^A$ because each respondent will be classified as either a master or a non-master of each attribute. The models are confirmatory because the attributes needed to respond to each item correctly are also specified a priori.

DCMs are typically used to measure knowledge at a finer grain size than IRT or CTT. For instance, an IRT-based analysis might estimate math ability in the domain of adding fractions while a DCM-based analysis could be used to diagnose the distinct skills needed to add fractions—e. g. adding whole numbers, finding common denominators, and changing improper fractions to mixed numbers. They provide three levels of feedback that can be used to inform instructional decisions: 1) the distribution of skill classes within the test population, 2) the frequency of mastery for each skill in the test population, and 3) the most probable skill profile for each student (George et al., 2016). The finer-grained information that DCMs offer (compared to IRT and CTT models) comes with a price. They typically require longer tests and more respondents to estimate. As with IRT models, different DCMs have different numbers of parameters which depend upon the number of attributes, the number of items, and the relationships that are specified between the attributes. More expansive models require greater amounts of data for estimation. Because the number of latent classes increases exponentially with the number of attributes, most applications of DCMs are limited to a maximum of six attributes (Rupp & Templin, 2008).

There are many DCMs for both dichotomous and polytomous response data. Rupp et al. (2010) provide a "taxonomy of DCMs" (p. 97) based on the types of response and latent variables (dichotomous or polytomous) and whether they are *compensatory* or *noncompensatory*. Compensatory DCMs are those for which the probability of answering an item correctly increases only when all attributes the item calls upon have been mastered. In noncompensatory DCMs, it is hypothesized that having some, but not all, of the attributes needed to answer an item correctly increases the probability of a correct response. Rupp et al.'s (2010) classification scheme allows models to appear in a single category or multiple categories. Next, I provide a selection of some of the models. The development of new models is a popular research topic. The examples that follow are a representation of the many models that are available.

Noncompensatory models that accommodate dichotomous response and latent variables include the *deterministic input, noisy-and-gate* (*DINA*) model (Junker & Sijtsma, 2001); the *higher order DINA* (*HO-DINA*) model (de la Torre & Douglas, 2004); the *multiple strategies DINA* (*MS-DINA*) model (de la Torre & Douglas, 2008); *noisy input deterministic-and-gate* (*NIDA*) model (Junker & Sijtsma, 2001); the *rule-space method* (*RSM*) (Tatsuoka, 2009); the *attribute hierarchy method* (*AHM*) (Gierl et al., 2007); three versions of the *reparameterized unified model* (*RUM*) (DiBello et al., 1995)—the *randomized effects RUM* (*RERUM*) and the *noncompensatory RUM* (*NC-RUM*) model in both a full and reduced version; *Bayesian inference networks* (*BINs*) (Sinharay & Almond, 2007); and *multiple classification latent class models* (*MCLCMs*) (Maris, 1999). Some of these models (DINA and NIDA) are designed specifically for dichotomous variables while others are more general. For instance, the RSM and the AHM can accommodate polytomous latent variables and the full and reduced NC-RUM models can accommodate polytomous response and latent variables. Compensatory models that

accommodate only dichotomous response and latent variables include the *deterministic inputs, noisy-or-gate* (*DINO*) model (Templin & Henson, 2006) and the *noisy inputs, deterministic-or-gate* (*NIDO*) model. Compensatory models that accommodate both dichotomous and polytomous response and predictor variables include the *compensatory RUM* (*C-RUM*) model; the *general diagnostic model* (*GDM*) (von Davier, 2005); the *loglinear cognitive diagnosis model* (*LCDM*) (Henson et al., 2009); the *generalized DINA* (*G-DINA*) model (de la Torre, 2011); and the *hierarchical GDM* (von Davier, 2007). Models differ in both their generality and their complexity. The most general models may be the full and reduced NC-RUM which can accommodate both dichotomous and polytomous variables for compensatory models, and BINs and MCLCMs which can be used to model any combination of dichotomous and polytomous variables and to create both compensatory and noncompensatory models. The more general models often require the estimation of a larger number of parameters and may require larger sample sizes and longer tests. More complex models may fit real data better than less complex models, but they are also less likely to converge during estimation and are more prone to overfitting (Rupp et al., 2010). The choice of a suitable model should consider these issues.

For the MAFA, it seems unlikely that possessing only one of multiple misconceptions would increase the probability of a "correct" answer. Because misconceptions are measured using true/false items, respondents are not forced to accept an answer they believe is partially correct (i.e. an item for which the respondent has only one of multiple misconceptions that are required) as they might with some multiple-choice items. Instead, they can choose "false". Therefore, this project uses a noncompensatory model for the measurement of misconceptions. The three core noncompensatory DCMs according to Rupp at al. (2010) are the DINA model, the NIDA model, and the NC-RUM model. These models differ in the ways that they account for

slipping—failing to answer an item correctly when one possesses the necessary attributes—and *guessing*—answering an item correctly although one does not possess the necessary attributes. The DINA model accounts for slipping and guessing at the item level, the NIDA model does so at the attribute level, and the NC-RUM model does so at both the item and attribute levels. These differing specifications affect the number of parameters that must be estimated for each model. In the DINA model, there is one slipping and one guessing parameter for each item regardless of the number of attributes measured. In the NIDA model, there is one slipping and one guessing parameter per attribute regardless of the number of items. In the NC-RUM model, there is one slipping related parameter for each item and one parameter related to both slipping and guessing for each item-required attribute combination. According to Rupp and Templin (2008), both the DINA and NIDA models are likely to converge for tests of 20-40 items and 4-6 attributes when sample sizes are as low as a few hundred while the NC-RUM model would require a much larger sample size.

Due to its parsimonious nature, the DINA is the most widely used core DCM (George et al., 2016). Among the noncompensatory models, the DINA model seemed best suited for the purpose of this research—to model student misconceptions with a test that has a small number of items and with a reasonable sample size—for two reasons. First, physics misconceptions may be context specific and are best modeled at the more context-specific item level as in the DINA model. Second, the sample size of 449 respondents was adequate to achieve model convergence. In the next section, I briefly describe other DCMs that have been proposed for measuring misconceptions and explain why they were unsuitable for this research.

Existing DCMs for Measuring Misconceptions

A search of the literature revealed three recent DCMs that have been proposed to measure misconceptions. All three were considered as possible models to use in this research. These models are the *Bug-DINO* (*Bug-diagnostic input noisy or gate*) model (Kuo et al., 2016), the *SISM* (*simultaneously identifying skills and misconceptions*) model (Kuo et al., 2018), and the *SICM* (*scaling individuals and classifying misconceptions*) model (Bradshaw and Templin, 2014). This section briefly describes each of the three models and outlines the reasons for creating a different model for this project.

The term "bug" in the Bug-DINO model (Kuo et al., 2016) refers to misconceptions. The model provides estimates of student misconceptions only. Correct answers are due to the absence of misconceptions and make no assumptions about the possession of skills, which are modeled separately. It is assumed that the possession of even one misconception related to an item will result in an incorrect response. Each item is scored as correct or incorrect, but distractors are not scored. Instead, it is assumed that an incorrect answer is caused either by possessing at least one of the misconceptions aligned with the item or by a slip. The probability that a student gives a correct answer to an item because he possesses none of the misconceptions aligned to the item is given by

$$P\left(X_{ij}=1\middle|\alpha_i\right)=(1-s_j)^{1-\xi_{ij}}(g_j)^{\xi_{ij}} \tag{9}$$

where α_i is a vector of student *i*'s misconceptions, s_j is the probability of slipping for item *j*, g_j is the probability of guessing for item *j*, and ξ_{ij} equals "1" if a person *i* possesses at least one of the misconceptions aligned to item *j* and "0" if person *i* possesses none of the misconceptions aligned to item *j*. The authors of the model apply it to computational multiple-choice items

(proportional reasoning in math). In this context, it is reasonable to assume that possessing any misconception would cause a student to answer an item incorrectly. For the MAFA, however, such an assumption is not reasonable. First, many physics students answer computational problems correctly despite possessing misconceptions (Mazur, 2009). Second, when measuring misconceptions using conceptual items, particularly true/false items, the possession of fewer than all of the required misconceptions is unlikely to cause a student to choose that answer. Endorsing a statement that relates to multiple misconceptions is likely to require the possession of all of them. For these reasons, the Bug-DINO model did not fit the design of this assessment.

In contrast to the Bug-DINO model, the SISM and SICM models estimate both knowledge and misconceptions. The SISM model essentially combines the DINA model for measuring knowledge with the Bug-DINO model for measuring misconceptions in a way that allows for the coexistence of knowledge and misconceptions (Kuo et al., 2018). Items are scored as correct or incorrect and each item is aligned with attributes that may be either knowledge, misconceptions, or both. Conceptually, the model assumes that the probability of a correct answer depends on how many skills (all or some) and misconceptions (some or none) a student has. The model estimates the probability of a correct response as

$$P\left(X_{ij}=1\middle|\boldsymbol{\alpha_i}\right)=h_j{}^{n_{ij}(1-\gamma_{ij})}\omega_j{}^{\eta_{ij}\gamma_{ij}}g_j{}^{(1-\eta_{ij})(1-\gamma_{ij})}\varepsilon_j{}^{(1-\eta_{ij})\gamma_{ij}} \tag{10}$$

where X_{ij} equals person *i*'s answer to item *j*, $\boldsymbol{\alpha_i}$ is a vector of person *i*'s attributes, η_{ij} equals "1" if person *i* has all required skills for item *j* and "0" if they do not, γ_{ij} equals 1 if person *i* has some of the misconceptions associated with item *j* and "0" if they do not, and the remaining parameters are the probabilities of a correct answer to item *j* for a person who has all of the skills and none of the misconceptions (h_j), all of the skills and only some of the misconceptions (ω_j),

some of the required skills and none of the misconceptions (g_j), or some of the skills and at least one of the misconceptions (ε_j). When the researchers used the model to estimate knowledge and misconceptions from an existing seven-item fractions test, a high rate of agreement was found between the results and classifications of human raters on the same data. The model had a higher agreement with human raters than the Bug-DINO model. Despite the high performance of the model on fractions data, there were two reasons to use a different model for the MAFA. First, it is preferable to measure knowledge of Newton's laws as a unidimensional construct because this is what the original FCI does. The practice of using a single score for FCI performance is well established in the physics community and has proven useful in many studies. Second, scoring only correct answers means that all misconceptions associated with an item will be linked to that item in the Q-matrix. For the MAFA this would result in most items aligning to most misconceptions. In this case, it is likely that most people would be assigned attribute profiles with either all or none of the misconceptions. The opportunity for a finer grained diagnosis of misconceptions would be lost.

The SICM model (Bradshaw & Templin, 2014) corrects for some of the aforementioned issues with the prior two models. It estimates ability as a continuous latent variable using IRT and misconceptions as an attribute profile using a DCM. It aligns misconceptions with the individual options for each item and requires that each option be scored. The endorsement of the correct option contributes to a higher knowledge score and the endorsement of an incorrect option increases the probability of having the misconceptions that are aligned with that option. In this way, it uses response data to give a more fine-grained view of misconceptions. However, it does not allow for the coexistence of knowledge and misconceptions. The complex specification of the model is beyond the scope of this paper. The authors tested the model using both

simulated data and a data set of 10,039 FCI responses. The simulation study indicated that relatively high numbers of items (30 or more) and responses (3000 or more) produced accurate classification rates for misconception profiles but did not produce accurate ability estimates. Analysis of the FCI data highlighted some of the problems with using the original FCI to diagnose misconceptions. First, the authors of the FCI list 31 misconceptions (Hestenes et al., 1992) which is far too many to model with existing DCMs. For this analysis, the researchers randomly chose to diagnose the first three misconceptions in the list. Second, many of the response options did not align with any of the measured misconceptions which created responses with only entries of "0" in the Q-matrix. This points to one of the differences between existing concept inventories and a CDA specifically designed to diagnose misconceptions. A CDA can be designed to specifically target a limited number of misconceptions and, thus, be a more efficient tool.

All three models described above are designed to elicit diagnostic information from traditional tests in which there is a correct answer and distractors which align with common misconceptions. This research took a different approach by redesigning the format of the assessment to one composed of separate sets of questions to measure knowledge and misconceptions. Such an assessment can use existing psychometric models to separately estimate ability and misconceptions. The psychometric analysis combined variables of the three models described above to achieve the desired results. Like the Bug-DINO model, it estimates knowledge and misconceptions separately. Like the SISM model it allows for the coexistence of knowledge and misconceptions. Like the SICM model, it models knowledge as a continuous variable using IRT and misconceptions as latent skills modeled with the noncompensatory DINA model. However, it uses separate sets of test items for each psychometric model and estimates

each separately. It was hoped that this approach would allow accurate estimation of ability and the most probable profile of misconceptions for each respondent using fewer items. In the next section, I describe the DINA model and how it was used for this assessment in greater detail.

The DINA Model

The DINA model estimates the probability that each respondent is in one of two classes for each item: 1) the group that has mastered all attributes which are relevant to the item, or 2) the group that has failed to master at least one of the attributes relevant to the item. The probability of a correct response is defined at the item level as the combined probability of either guessing or not slipping:

$$\pi_{jc} = (1 - s_j)^{\xi_{jc}} {g_j}^{1-\xi_{jc}} \tag{11}$$

where π_{ic} is the probability of a correct response to item *j* for a respondent in latent class *C*, s_j is the probability of slipping for item *j*, g_j is the probability of guessing for item *j*. The guessing parameter (*g*) is the probability of answering correctly even though one does not possess all required attributes and the slipping parameter (*s*) is the probability of failing to answer correctly even though one possess all required attributes. The final variable, $\xi_{jc,}$ is a latent variable that relates the skills needed to correctly answer item *j* to the skills possessed by respondents in latent class *c*. This is the *deterministic-input* part of the DINA model and is calculated as follows:

$$\xi_{jc} = \prod_{\alpha=1}^{A} \alpha_{ca}^{q_{ja}} \tag{12}$$

where α_{ca} equals 1 if respondents in class *c* have mastered attribute *a* and 0 if they have not, and q_{ja} is the entry in the Q-matrix for item *j* and attribute *a* which equals 1 if item *j* requires the use

of attribute *a* and 0 if it does not. This conjunctive condensation rule results in $\xi_{jc} = 1$ for items and classes that require/indicate mastery of the same attributes and $\xi_{jc} = 0$ for classes in which one or more attributes relevant to item *j* have not been mastered. Working backwards, it becomes evident that the probability of a correct response to item *j*, π_{jc}, will equal $(1 - s_j)$ for respondents who have all attributes needed for item *j* and g_j for respondents who are missing at least one attribute needed for item *j*. Because the Q-matrix and attribute profiles are specified a priori, estimating values for this part of the model focuses on the slipping and guessing parameters for each item.

Not only are the slipping and guessing parameters unknown. The latent class to which each respondent belongs is also unknown. The probability that the vector of person *i*'s responses belongs to an individual in latent class *c* is given by

$$P(\boldsymbol{X}_i|\boldsymbol{\alpha}_c, \boldsymbol{g}, \boldsymbol{s}) = \prod_{j=1}^{J} P\big(X_{ij} = 1\big|\alpha_c, g_j, s_j\big)^{X_{ij}}[1 - P\big(X_{ij} = 1\big|\alpha_c, g_j, s_j\big)]^{1-X_{ij}} \qquad (13)$$

where $\boldsymbol{X}_i$ is the vector of person *i*'s responses, $\boldsymbol{\alpha_c}$ is the vector indicating the attribute profile for individuals in latent class *c*, $\boldsymbol{g}$ is the vector of guessing parameters for the items on the test, and $\boldsymbol{s}$ is the vector of slipping parameters for items on the test. The addition of the *j* subscript to any variable indicates that it is the *j*th member of the vector—the value for item *j*. Estimating these values is fairly straightforward using estimated item parameters and allows the assignment of the most probable attribute profile to each respondent.

Parameter Estimation for the DINA Model

Estimating the item parameters for the DINA model involves calculating two values, *s* and *g*, for each item. Because the DINA model models slipping and guessing at the item level,

the number of parameters that must be estimated is independent of the number of attributes which are measured and depends only on the number of items. This contrasts with other non-compensatory DCMs which model aberrant responses at the attribute level (NIDA) or at both the item and attribute level (NC-RUM). A potential problem with the use of this model for true-false items is that it requires $(1-s_i) > g_i$. The probability of guessing at random for true-false items is 0.5 which could be problematic. However, because the true-false items on this assessment address common misconceptions in everyday situations, they are likely to be either firmly held or absent. It is assumed less likely that students would guess on these items than on the knowledge items.

The marginal log-likelihood function which was used to find the most probable ability in IRT is used here to find the most probable item parameters. The marginal log-likelihood for the DINA model is given by

$$\ln L(\boldsymbol{g}, \boldsymbol{s}) = \sum_{i=1}^{I} \ln(\boldsymbol{X}_i, \boldsymbol{g}, \boldsymbol{s}) = \sum_{i=1}^{I} \ln[P(\boldsymbol{X}_i|\boldsymbol{\alpha}_c, \boldsymbol{g}, \boldsymbol{s}) \cdot P(\boldsymbol{\alpha}_c)] \tag{14}$$

where all variables are defined in the prior section of this chapter. The CDM software package implements the MMLE estimation using the EM algorithm. This is an iterative process whereby the expected number of students in each attribute profile is estimated, these are used to estimate the item parameters which are then used to estimate the distribution of attribute profiles, and so on until the model reaches convergence—the point at which the change in estimates between iterations reaches some lower threshold. Next, I describe this process in greater detail based on George et al.'s article (2016).

The notation in this part of the paper follows that of George et al. (2016) except that I use *c* rather than *l* to refer to latent classes. The first E-step of the algorithm requires some estimate of the slipping and guessing parameters to begin. These estimates are used with Equation 13 and Bayes' theorem to calculate the individual posterior distribution $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i, \boldsymbol{g}, \boldsymbol{s})$. This is used to calculate the expected number of students with each attribute profile based on item *j* and the expected number of students with each attribute profile who responded to item *j* correctly. This completes the E-step. In the M-step, these expected values across all attribute profiles and the resulting values are used to calculate the following values for each item *j*:

$T_j^{(0)}$ the expected number of students who lack at least one required attribute for item *j*

$R_j^{(0)}$ the expected number of students who lack one attribute, but answered the item correctly

$T_j^{(1)}$ the expected number of students who possess all attributes for item *j*

$R_j^{(1)}$the expected number of students who possess all attributes and answered the item correctly.

The item parameters are then estimated again as

$$\hat{g}_j = \frac{R_j^{(0)}}{T_j^{(0)}} \tag{15}$$

and

$$\hat{s}_j = \frac{R_j^{(1)}}{T_j^{(1)}} \tag{16}$$

Next, the expected number of students with each attribute profile is calculated and used to update the distribution of skill profiles, $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i, \boldsymbol{g}, \boldsymbol{s})$. This concludes the M step. The E- and M-steps are repeated until the model reaches convergence or a specified number of iterations has completed. At this point we have item parameter estimates which can be used to estimate individual attribute profiles.

The CDM Package can estimate individual attribute profiles using three different methods: maximum a posteriori (MAP) classification, maximum likelihood estimation (MLE), and expected a posteriori (EAP) classification. Maximum likelihood estimation does not provide attribute profiles for respondents who answer all items correctly or all items incorrectly whereas MAP and EAP estimation do. Both MAP and EAP estimation apply a prior distribution for latent class membership within the population. Then the probability of each attribute profile/latent class (MAP) or probability of individual attribute mastery (EAP) is calculated for each response pattern. These are used to place each respondent into the most likely attribute profile.

Model Fit for the DINA Model

In assessing model fit for DCMs, it can be convenient to think of three general areas: item parameters, person classification, and overall model fit. Overall model fit of the DINA model to the data can be assessed using the AIC and the BIC which were described in an earlier section of this chapter. For these fit statistics, a lower value indicates better fit. In addition, the CDM package (Robitzsch et al., 2020) provides a chi-square statistic that compares the observed response probabilities to the predicted response probabilities. Other fit statistics in the package that can be utilized are the root mean-square error of approximation (RMSEA) for assessing item fit, an item discrimination index that tests the constraint that $g_j < (1 - s_j)$, and a simulation for checking the accuracy of attribute profile classification. These will be employed to create an

inclusive picture of model fit. Possible sources of misfit include the incorrect choice of model type (compensatory or noncompensatory), inaccurate specification of the Q-matrix, a poor choice of model restrictions (i.e. restrictions across parameters or prior distributions), and a heterogeneous population (Rupp & Templin, 2008).

Summary

This chapter has provided a background on topics within the domains of science education and educational measurement that are relevant to the research project. Here I provide a brief summary of the major points. Many students enter formal science instruction with conceptions about science that are incorrect and may inhibit the development of scientifically accurate understanding of phenomena. Often, science instruction fails to correct this problem. Concept inventories have been developed as time-efficient tools for measuring the extent to which students are able to think scientifically about given topics such as mechanics, climate change, and astronomy. These inventories usually take the form of multiple-choice tests in which the distractors represent common misconceptions about the subject, but they are not designed to diagnose the presence of specific misconceptions.

Using DCMs to diagnose student misconceptions has been suggested. A few DCMs have been developed specifically for this purpose. They have been tested with simulated data and even retrofitted with data from an existing concept inventory, the FCI. Retrofitted data is unlikely to fit as well or to provide as accurate a classification of respondents as are data from an assessment that is designed to be fit with a DCM, a CDA. In addition, two of the existing models do not allow for the coexistence of knowledge and misconceptions on an item, and the third model is likely to classify most examinees on the MAFA into two classes, those who possess all

misconceptions and those who possess no misconceptions. Therefore, this research proposes to develop and test a CDA based on the DINA model. In addition to fitting the structure of the proposed items, the DINA model has the advantage of being the most parsimonious of the DCMs. This means that it can be estimated well with fewer items and respondents. Because this assessment is intended to be used formatively, fewer items are appropriate. The final sections of the chapter provide an introduction to IRT and DCMs. They explain the statistical models involved and the methods of parameter estimation along with measures of model fit. The next two chapters consist of two independent manuscripts. The first manuscript (shown in Chapter 3) presents a test format for CDAs and describes the test creation process for the MAFA and the second manuscript (shown in Chapter 4) presents describes the relationship between knowledge and misconceptions about Newton's first and second laws as measured by the MAFA.

Chapter 3

Development of a Diagnostic Cognitive Assessment for Measuring Misconceptions About Force

Abstract

A new assessment format for simultaneously measuring knowledge and misconceptions using item response theory and diagnostic cognitive modeling was proposed. Application of the proposed model was illustrated by the development of an online assessment for measuring knowledge and misconceptions about Newton's first and second laws--the Misconceptions About Force Assessment. The assessment was developed in four phases. First, items were created by the author based on prior research about student misconceptions of force and motion. Experienced high school and university instructors reviewed the items and misconceptions for content and alignment. For Phases two through four, all participants were undergraduate university students from public four-year institutions who had completed no more than two semesters of university level physics courses. In Phase 2, think-alouds were conducted with four students. In Phase 3, a pilot test was conducted with 100 students and the results were used to assess item quality. In Phase 4, a field test was conducted with 349 students. Data from Phases 3 and 4 were combined and fit to item response models and a cognitive diagnostic model. Model fit was acceptable for both models. Suggestions for further research include testing a shorter version of the assessment with a larger and more motivated group of participants, comparing model fit for additional combinations of items, and applying the test format to develop other concept inventories. The research demonstrates a process for applying diagnostic cognitive

models to create concept inventories that can provide simultaneous measures of student knowledge and profiles of student misconceptions.

Introduction

Students who perform well on typical classroom assessments do not always understand what had been taught. A classic example of this is shown in the film "A Private Universe" in which engineers who have just graduated from MIT are unable complete a simple circuit to light a small bulb with a battery and a single wire (Schneps & Sadler, 1988). Although the newly minted engineers must have successfully completed many difficult assessments to earn their degrees, they lack the conceptual understanding of electricity that is needed to solve this simple, but novel, problem. Similar disconnects have been confirmed by other researchers in science education (Diakidoy & Iordanou, 2003; Mazur, 2009) and reading and math education (Eckert et al., 2006).

In education, most models of learning are based on the belief that students construct new knowledge by connecting new ideas and experiences to their existing conceptions (Jones & Bradjer-Araje, 2002; Lucariello & Naff, n.d.; Matthews, 1998; National Research Council [NRC], 1997; Phillips, 1995). The process of learning science is complicated by students' preexisting knowledge and beliefs that they use to explain the world around them because much of that knowledge is incorrect and resistant to change (Brown & Hammer, 2013; Duit & Treagust, 2003; Vosniadou & Skopeliti, 2014). Because student's existing conceptions about science are often incorrect, the new knowledge that they construct is often incorrect as well. There are many ways that this can occur. For instance, a student may combine new knowledge with incorrect existing conceptions to create a partially correct model or the student may distort

the new information to create a completely incorrect model. In comparison to science experts, students tend to use large numbers of unrelated, personal theories and beliefs to explain natural phenomena. One aim of science education is to help students learn to think more like science experts. Progression toward this status requires students to undergo conceptual change in which they adapt their existing conceptions and beliefs (diSessa & Sherrin, 1998; Duit & Treagust, 2003).

Learning science is a messy process that occurs over time. Researchers employ different terms to describe the extent to which student's knowledge at a given time is correct. For instance, *misconceptions*, (Halloun & Hestenes, 1985a; McCloskey et al., 1983) refer to student knowledge and beliefs that are incorrect whether they are preexisting or have been developed though instruction. Other researchers use the term *preconceptions* (Clement, 1982; Halloun & Hestenes, 1985b) or *alternative* conceptions (Chi et al., 1981; Viennot, 1985) to refer to both correct and incorrect preexisting ideas that students bring to instruction. This terminology is used to emphasize that student's existing knowledge—both correct and incorrect--may be used by teachers to increase the effectiveness of instruction. Student's preconceptions can be identified through multiple methods such as informational interviews, written assessments, and observing students working with equipment to complete tasks. There is a rich body of research on students' misconceptions in science, especially in physics (Duit, 1993).

Researchers who study conceptual change vary in the extent to which they believe that students' preconceptions are organized and called upon consistently. For instance, Vosniadou and Skopeliti (2014) write about framework theories which are loosely structured sets of related concepts based on everyday interactions and experiences that reside within ontological beliefs. Framework theories are activated fairly consistently to explain new experiences They give the

example of "naive physics" as one such framework theory. In contrast, diSessa (1993) writes about phenomenological primitives (p prims) which are small pieces of knowledge or beliefs that have been developed from everyday experience. A single p prim tends to be limited to explaining a small number of phenomena. In this model of novice knowledge, individuals develop and activate many p prims to explain their experiences and observations. Smith et al. (1993) use the term conceptions to refer to student ideas that differ from those of experts, but that are called upon consistently to make sense of natural phenomena and strongly affect how students learn science. To develop accurate scientific knowledge, students must confront their existing misconceptions and reconstruct their mental models (NRC, 1997). However, the extent to which students are open to adapting their preconceptions even when they are incorrect varies (Harrison & Treagust, 2001; Vosniadou, 2014). Teachers can help students achieve conceptual change by providing feedback about their misconceptions (Hattie, 2015), but first the misconceptions must be identified.

When common misconceptions in an area have been identified, concept inventories can be used to measure the prevalence of the misconceptions. Concept inventories are multiple-choice tests in which the distractors correspond to common misconceptions. Concept inventories are sometimes used to measure the effect of classroom instruction on learning. The test is administered before and after instruction and the mean change in class score (calculated as the normalized gain) is used to compare student performance under different conditions (Hake, 1998; LoPresto & Murrell, 2011; Thornton, et al., 2009; Williamson, et al., 2016; Yeo & Zadnick, 2001). An example of a test which is often used this way is the Force Concept Inventory (FCI) (Hestenes, et al., 1992) which is the most widely used concept inventory in physics education (Smith & Tanner, 2010). Some of the many concept inventories that have been

developed in various science disciplines are shown in Table 3.1. Although the distractors in concept inventories are based on common misconceptions, they are not typically scored to measure which misconceptions students have. Instead, an overall score is calculated based on the number of correct answers and it is assumed that the higher the total score, the fewer misconceptions a student has.

Table 3.1

Examples of Concept Inventories and Scoring Methods by Discipline

Discipline	Name and Reference	Scoring
Physics	Force Concept Inventory (Hestenes et al., 1992)	CTT
	Mechanics Baseline Inventory (Hestenes & Wells, 1992)	CTT
	Force and Motion Conceptual Evaluation (Thornton & Sokoloff, 1998)	CTT
	Brief Electricity and Magnetism Assessment (Ding et al., 2006)	CTT
	Thermal Concept Evaluation (Yeo & Zadnick, 2001)	CTT
	Newtonian Gravity Concept Inventory (Williamson, 2013)	CTT & IRT
Chemistry	Chemistry Concept Inventory (Pavelich et al., 2004)	CTT
	Quantum Chemistry Concept Inventory (Dick-Perez et al., 2016)	CTT
Astronomy	Light and Spectroscopy Concept Inventory (Bardar et al., 2006)	CTT
	Astronomy and Space Science Concept Inventory (Sadler et al., 2010)	CTT
	Astronomical Misconceptions Survey (LoPresto & Murrell, 2011)	CTT
Biology	Conceptual Inventory of Natural Selection (Anderson et al., 2002)	CTT
	Biology Concept Inventory (Klymkowsky et al., 2003)	CTT
Geoscience	Climate Change Inventory (Jarrett et al., 2012)	CTT
	Geoscience Concept Inventory (Libarkin & Anderson, 2005)	IRT

As shown in Table 3.1, most concept inventories are scored according to a measurement model belonging to Classical Test Theory (CTT), but a few have been scored using Item Response Theory (IRT). These are two of three common approaches to scoring large-scale

assessments. The third is Latent Class Analysis (LCA). In CTT, assessment scores are based on some version of number of items answered correctly. CTT can be used to analyze responses to assessments that measure a single latent construct even for a small number of responses. IRT is a family of probabilistic latent construct models—some unidimensional and some not-- that relate the assessment scores to the probability of answering items of different difficulties correctly. IRT offers some advantages over CTT for unidimensional assessments, but larger numbers of responses are required to estimate the models. LCA uses probabilistic models that measure the presence or absence of multiple latent constructs (sometimes called *attributes* or *skills*) based on response patterns. One group of relatively recent restricted latent class models are diagnostic cognitive models (DCMs). These models estimate the presence of a set of multiple attributes/skills (e.g. finding a common denominator for fractions, fear of strangers, etc.) based on responses to selected response items. They are useful for modeling responses to assessments which measure multiple skills or attributes. The three approaches to analyzing assessment responses—CTT, IRT, and DCMs—are built on different sets of assumptions and can be used to provide different information about student performance.

Measurement models based on IRT have some advantages over simpler CTT models. First, IRT-based tests can provide unbiased estimates of person abilities with fewer items than CTT-based tests. In CTT, a person's true score--the score they would attain if they were to answer every possible item about the tested topic an infinite number of times-- is assumed to be the difference between their observed score on a test and an error term. While a person's true score cannot be directly measured, it can be estimated by the observed score. Therefore, the smaller the error term, the closer the observed score is to the true score. Because the error term for each test item is random, the longer the test, the closer the error term gets to zero and the

more accurate the observed score. Second, in IRT item difficulty and item discrimination are independent of who takes the test. In CTT item difficulty is defined as the mean score for an item. For dichotomous items this is the percent of respondents who answer an item correctly. The same item may have a low difficulty with students of high ability and a high difficulty with students of low ability. This is at odds with the idea of latent constructs which "occupy a latent space that can be quantified along a hypothesized infinity continuum from (-∞, ∞)" (Osterlind, 2010, p. 273). Item response models solve this problem by measuring item difficulty and person ability on the same scale. In the two of the most used IRT models (the 1-PL and 2-PL models), the difficulty of a dichotomous item is defined as the ability at which an individual has a 50% chance of answering the item correctly. This is halfway between the probability of a correct response for a person with ability of -∞--a probability of 0.0--and for a person with an ability of +∞--a probability of 1.0. Third, IRT allows the standard error of measurement (SEM)—the standard deviation of the measurement error on a test--to vary across ability levels. In CTT, SEM is constant across all levels of ability. This may not be accurate—especially for very low or very high abilities. Finally, IRT approaches reliability with the test information function—a measure of the abilities for which the test gives the most precise estimates. In CTT reliability is described at the test level and the assumptions required to calculate a measure of reliability (some version of parallel test forms) may be difficult to meet.

The three most used IRT models for dichotomous responses are the one-parameter logistic (1-PL), two-parameter logistic (2-PL), and three-parameter logistic (3-PL) models (Osterlind, 2010). Each model provides an estimate of person ability along an arbitrary scale that is also used to measure item difficulty. It is assumed that the ability being measured is approximately normally distributed within the population and IRT software generally scales

ability along the standard normal distribution, N(0,1). Osterlind (2010) lists three additional important assumptions for the 1-PL and 2-PL models. First, the test should be "unidimensional" meaning that all test items should measure the same, single latent construct. Second, test items are locally independent. This means that responses to all items should depend only on the latent construct being measured. Unusually high correlations between item responses after controlling for person abilities (i.e. correlations between residuals) may indicate that this is not the case. Such correlations may occur for items that share information such as a reading passage or diagram. A third important assumption is that the model fits the data reasonably well. Osterlind refers to a fourth assumption, which applies to assessments that measure academic knowledge. This is the idea that respondents apply their ability to every item. There is no part of the statistical model that accounts for respondents using less than their maximum ability on any item and certainly no easy way to test that they have.

Mathematically, the 1- and 2-PL models relate the probability of a correct response on item j to person ability (θ), item difficulty (β), and item discrimination (α) as:

$$p\left(x_j = 1 \middle| \theta, \alpha_j, \beta_j\right) = \frac{e^{\alpha_j(\theta-\beta_j)}}{1 + e^{\alpha_j(\theta-\beta_j)}} \tag{4}$$

where $x_j = 1$ indicates a correct response to item j and $x_j = 0$ indicates an incorrect response. For the 1-PL model, item discrimination is held constant across items and in the 2-PL model item discrimination is allowed to vary. For both models, when person ability and item difficulty are equal, the probability of a correct response is 0.5. Item discrimination measures how quickly the probability of a correct answer changes as item difficulty changes. The higher the

discrimination, the quicker the change and, therefore, the better the item discriminates between persons with different abilities.

Important factors to consider when choosing an appropriate model include test length and sample size. As model parameters are added, some combination of a larger sample size and/or a longer test is required for estimation. Despite the abundance of research concerning the effects of these factors on model fit and parameter estimation, there are no exact guidelines. However, recommendations suggest that for the 2-PL model, a 10-item test should have a minimum sample size of 750 (Alpir & Duygu, 2017) while a 20-item test should have a sample size of about 500 (Alpir & Duygu, 2017; de Ayala, 2009). Yen and Fitzpatrick (2006) note that shorter tests and smaller sample sizes may be used in low stakes applications such as during field testing.

Diagnostic Cognitive Models

Diagnostic assessments in education may provide information about student's attributes which teachers can use to individualize instruction. Diagnostic cognitive models (DCMs) provide one way to interpret responses to diagnostic assessments. DCMs are confirmatory multidimensional restricted latent-class models in which each latent class is a profile showing whether a person possesses each of a set of *attributes*. Attributes are latent states that are needed to respond correctly to the items on the assessment. Single items may require one or more attributes to answer correctly. The number of latent classes is restricted by the number of attributes. These are specified a priori. If the number of attributes specified in the assessment is A, then the number of possible latent classes (*N*) is given by $N = 2^A$ because each respondent will be classified as either a master or a non-master of each attribute. The models are confirmatory because the attributes needed to respond to each item correctly are also specified a priori. Attributes are also referred to as *skills*.

DCMs are typically used to measure knowledge at a finer grain size than IRT or CTT. For instance, an IRT-based analysis of test responses might provide an estimate of math ability in the domain of adding fractions while a DCM-based analysis could be used to diagnose the distinct skills needed to add fractions (e.g. adding whole numbers, finding common denominators, and changing improper fractions to mixed numbers). They provide three levels of feedback that can be used to inform instructional decisions: 1) the distribution of skill classes within the test population, 2) the frequency of mastery for each skill in the test population, and 3) the most probable skill profile for each student (George et al., 2016). The finer-grained information that DCMs offer (compared to IRT and CTT models) comes with a price. They typically require longer tests and more respondents to estimate. As with IRT models, different DCMs have different numbers of parameters which depend upon the number of attributes, the number of items, and the relationships that are specified between the attributes. More expansive models require greater amounts of data for estimation. Because the number of latent classes increases exponentially with the number of attributes, most applications of DCMs are limited to a maximum of six attributes (Rupp & Templin, 2008).

Many different DCMs have been developed and evaluated. Due to its parsimonious nature, the *deterministic inputs noisy-and-gate* (*DINA*) model (Junker & Sijtsma, 2001) is the most widely used core DCM (George et al., 2016). The DINA model accounts for the probability of a correct response at the item level. Two parameters are estimated for each item, the probability of slipping--answering incorrectly when all needed skills are possessed--and the probability of guessing—a correct answer by a person who does not possess the skills needed for the item. The DINA model is a *noncompensatory* DCM-- the probability of a correct response depends on possessing all the skills required to answer an item. If a person is missing even one

of the skills needed to answer an item correctly, the probability that they will choose the correct response is the same as that for a person who possesses none of the needed skills. Possessing some of the needed skills does not compensate for missing any of them. In contrast, for compensatory DCMs the greater the number of required skills a person possesses, the greater the probability of a correct response. A person who possesses some of the skills has a greater probability of answering correctly compared to a person who possesses none of the skills and a smaller probability than a person who possesses all the required skills. Rupp et al. (2010) list two other core noncompensatory DCMs which model slipping and guessing differently. The *noisy input deterministic-and-gate* (*NIDA*) model (Junker & Sijtsma, 2001) accounts for aberrant responses at the attribute level and the *noncompensatory reparameterized unified model* (*NC-RUM*) (DiBello et al., 1995) accounts for them at both the item and attribute level.

The DINA model estimates the probability that each respondent is in one of two classes for each item: 1) the group that has mastered all attributes which are relevant to the item, or 2) the group that has failed to master at least one of the attributes relevant to the item. The probability of a correct response is defined at the item level as the combined probability of either guessing or not slipping:

$$P(X_{jc} = 1) = \left(1 - s_j\right)^{\xi_{jc}} {g_j}^{1-\xi_{jc}} \tag{13}$$

where $P(X_{jc} = 1)$ is the probability of a correct response to item j for a respondent in latent class c, s_j is the probability of slipping for item j, g_j is the probability of guessing for item j. The guessing parameter (g) is the probability of answering correctly even though one does not possess all required attributes and the slipping parameter (s) is the probability of failing to answer correctly even though one possess all required attributes. The final variable, $\xi_{jc,}$ is a latent

variable that relates the skills needed to correctly answer item *j* to the skills possessed by respondents in latent class *c*. This is the *deterministic-input* part of the DINA model and is calculated as follows:

$$\xi_{jc} = \prod_{\alpha=1}^{A} \alpha_{ca}^{q_{ja}} \tag{14}$$

where α_{ca} equals 1 if respondents in class *c* have mastered attribute *a* and 0 if they have not, and q_{ja} is the entry in the Q-matrix for item *j* and attribute *a* which equals 1 if item *j* requires the use of attribute *a* and 0 if it does not. The Q-matrix indicates which attributes are required to correctly respond to each item and is usually specified by content experts. This conjunctive condensation rule results in $\xi_{jc} = 1$ for items and classes that require/indicate mastery of the same attributes and $\xi_{jc} = 0$ for classes in which one or more attributes relevant to item *j* have not been mastered. The probability of a correct response to item *j*, π_{jc}, will equal $(1 - s_j)$ for respondents who have all attributes needed for item *j* and g_j for respondents who are missing at least one attribute needed for item *j*. Because the Q-matrix and attribute profiles are specified a priori, estimating values for this part of the model focuses on the slipping and guessing parameters for each item.

Not only are the slipping and guessing parameters unknown. The latent class to which each respondent belongs is also unknown. The probability that the vector of person *i*'s responses belongs to an individual in latent class *c* is given by

$$P(\boldsymbol{X}_i|\boldsymbol{\alpha}_c, \boldsymbol{g}, \boldsymbol{s}) = \prod_{j=1}^{J} P\left(X_{ij} = 1 \middle| \alpha_c, g_j, s_j\right)^{X_{ij}} [1 - P\left(X_{ij} = 1 \middle| \alpha_c, g_j, s_j\right)]^{1-X_{ij}} \tag{15}$$

where $\boldsymbol{X}_i$ is the vector of person *i*'s responses, $\boldsymbol{\alpha_c}$ is the vector indicating the attribute profile for individuals in latent class *c*, $\boldsymbol{g}$ is the vector of guessing parameters for the items on the test, and $\boldsymbol{s}$ is the vector of slipping parameters for items on the test. The addition of the *j* subscript to any variable indicates that it is the *j*th member of the vector—the value for item *j*. Estimating these values is fairly straightforward using estimated item parameters and allows the assignment of the most probable attribute profile to each respondent.

Estimating the item parameters for the DINA model involves calculating two values, *s* and *g*, for each item. Because the DINA model models slipping and guessing at the item level, the number of parameters that must be estimated is independent of the number of attributes which are measured and depends only on the number of items. The model requires that the probability of a correct response for someone who possesses all necessary skills is greater than the probability of guessing-- $(1-s_i) > g_i$. The slipping and guessing parameters can be estimated alongside the person classifications in an iterative manner using marginal maximum likelihood estimation (MMLE). First the expected number of people in each attribute class is calculated from a theoretical ability distribution. These are used to estimate the item parameters which are used to again estimate the number of people in each attribute file and so on. The process stops when the model converges. Convergence occurs when the difference between values in successive iterations reaches a specified threshold.

It has been suggested that DCMs could be used to measure students' misconceptions and three models have been developed for this purpose (Bradshaw & Templin, 2014; Kuo et al., 2016; Kuo et al., 2018). In each of the models, skills are replaced by misconceptions and a "correct" answer is the answer that would be chosen by a student who possesses the misconception. Instead of placing each respondent in the most likely attribute profile, the models

place each respondent in the most likely profile of misconceptions. Performance of the three models has been evaluated through simulation studies and/or by retrofitting data from existing assessments to them. The researcher found no CDAs that have been developed specifically for the purpose of assessing misconceptions. There are few assessments that have been designed using DCMs. Development of such assessments—sometimes called cognitive diagnostic assessments (CDAs)--requires expertise in both psychometrics and subject-matter knowledge literature (de la Torre, 2009). This is probably one reason for the small number of CDAs based on DCMs in the literature. Retrofitting models to existing data may produce questionable model and item fit and a high rate of examinee misclassification due to the violation of underlying assumptions (de la Torre & Minchen, 2014; Lee et al., 2012; Rupp & Templin, 2008). Development of CDAs in general and CDAs to measure misconceptions can contribute to the existing body of knowledge about DCMs.

Research Questions

The purpose of this research was to test the efficacy of a proposed test format for cognitive diagnostic assessments that measure knowledge and misconceptions. A new assessment about Newton's first and second laws of motion, the MAFA, was developed and evaluated using the test format. Hence, this is a proof-of-concept study. This paper addresses two specific questions:

1. How well do the specified measurement models (item response theory and deterministic input noisy-and-gate) fit responses to the MAFA?
2. How do responses on the MAFA compare to responses to Force Concept Inventory items which measure the same knowledge and misconceptions?

The next section of this paper describes the process of item development and the four phases of item analysis.

Method

Item Development

Test Specifications

The test specifications for the MAFA were created by the researcher and informed by prior research on students' misconceptions in physics. According to the *Standards for Educational and Psychological Testing*, test specifications include detailed statements "about content, format, test length, psychometric characteristics of the items and test, delivery mode, administration, scoring, and score reporting" (p.76) as well as the test's purpose and intended uses (American Educational Research Council [AERA] et al., 2014). The purpose of the MAFA is to diagnose students' misconceptions about and measure students' mastery of Newton's first and second laws of motion. The MAFA is composed of two sets of related items—*knowledge items* and *reason items*. Knowledge items are multiple-choice questions that ask about the numbers and directions of forces acting on objects and the types of motion that result from these forces. Responses to knowledge items are used to measure students' conceptual reasoning about force and motion within the context of Newton's first and second laws. They are scored using IRT. There are 18 knowledge items in the final version of the test. Reason items are true/false questions that ask about the reasons for answers to the knowledge items. Responses to the reason items are used to predict the probability that students possess any combination of six misconceptions about force and motion. Reason items are scored using a DCM which results in a profile of misconceptions for each respondent. Each knowledge item has between one and four

reason items associated with it. The format of the test is meant to mimic an interview in which students are asked to predict what will happen in a physical situation and then to explain the reasons for their answer. The test presents a single knowledge item at a time followed by the associated reason items. The test is administered online and it contains some item sets that are adaptive—either the reason items or answer choices to the knowledge item presented depend on the response that was chosen to the prior item. A sample item that is not about Newton's laws is given in Figure 3.1 to illustrate the test format. (This same item was used as a practice item during test administration). Results from the MAFA are intended to be used in classroom level

Figure 3.1

Sample Test Item

Knowledge Item	Choose the single best answer:
Item S	When a gas-filled balloon is placed in the freezer so that its temperature decreases, what happens to the volume of the air inside the balloon? --It decreases. --It stays the same. --It increases.
Reason Items	Indicate whether each statement about the gas molecules in the balloon is true or false:
Item S.1:	The gas molecules in the balloon get smaller when it is in the freezer. --True --False
Item S.2:	The space between the gas molecules decreases when the balloon is in the freezer. --True --False
Item S.3:	The space between the gas molecules increases when the balloon is in the freezer. --True --False
Item S.4:	The gas molecules stay the same size when the balloon is in the freezer. --True --False

formative assessment for both teachers and students. It is not intended to be used to assign grades or to make decisions about teacher performance.

The Test Domain

Because the MAFA measures misconceptions in addition to knowledge, the test domain needed to define both areas. There is a rich, research-based body of knowledge about student misconceptions in physics and the misconceptions included in the test domain were identified by reviewing the existing research. Figure 3.2 shows the content domain of the test including the parts of Newton's first and second laws that are assessed by the knowledge items and the final list of six misconceptions that are assessed by the reason items. For each misconception, citations for the original research in which they were identified are provided.

Many of the items on the FCI are based on the same research, therefore there are similarities between some of the items on this assessment and FCI items. One important difference is that the FCI measures conceptual knowledge of kinematics and all three of Newton's laws of motion while the test domain for the MAFA is limited to Newton's first and second laws. This contraction of the test domain is due to a combination of statistical and practical limitations. The greater the content domain, the greater the number of possible misconceptions there are to be diagnosed. This would require both a longer test and more participants to estimate the statistical models. Limiting the assessment to two laws and six misconceptions allowed the researcher to describe the misconceptions at a grain size that could be useful for instruction, to keep the test short, and to estimate the model with a reasonable amount of data while still demonstrating the development of a CDA to measure knowledge and misconceptions.

Figure 3.2

Content Domain for MAFA

Domain for Knowledge Items		
First Law	1.a.	If there are no outside forces acting on an object, it will continue in its state of motion—either at rest or in a straight line at a constant speed.
	1.b.	Objects that are either speeding up or slowing down have a non-zero net force acting on them.
	1.c.	Objects that are moving along a curved path have a non-zero net force acting on them with a component that is perpendicular to the line of motion.
Second Law	2.a.	Objects that are speeding up have a non-zero net force acting on them in the same direction they are moving.
	2.b.	Objects that are slowing down have a non-zero net force acting on them opposite the direction in which they are moving.
	2.c.	The bigger the net force acting on an object, the greater its acceleration.
	2.d.	Objects that are moving in a circle at a constant speed have a non-zero net force acting on them perpendicular to the direction in which they are moving/directed toward the center of the circle.

Domain for Reason Items		
Number	Misconception	Original Source and Physical Situations
1	When an object is moving in a given direction, there must be a force acting in that direction.	Clement, 1982 Rocket, Coin toss, & pendulum McDermott, 1984—hockey puck and air blast
2	An object moving in a curved path has an outward force acting on it.	Halloun & Hestenes, 1985b Unlu & Gok, 2007
3	A constant force causes an object to move with a constant velocity/ an object's velocity is proportional to the magnitude of applied force/changes in speed are caused by changes in the magnitude of applied force.	Wenning, 2008 Viennot, 1979 (in McDermott, 1984) Champagne et al., 1980 (in McDermott, 1984) Clement, 1982--Rocket, Coin toss, & pendulum
4	The force of gravity pulls on an object only when it is falling downward.	Wenning, 2008 Clement, 1982—Coin toss
5	An object that is moving in a curved path will continue to move in a curved path after the removal of the centripetal force.	McCloskey et al., 1980—Ball on string as seen from above, ball shot out of curved tube
6	An inanimate or passive object cannot exert a force on a second object because inanimate objects cannot push back.	Clement, 1998 (as cited in Cummings et al., 2004) Minstrell, 1982—Book on table, book on hand, book hanging from spring

In addition to content knowledge and misconceptions, a third aspect of the test domain was physical situation (e.g. a coin tossed into the air or a rocket moving through space). While physics experts tend to explain diverse physical situations using a small set of rules (e.g. Newton's laws of motion), novices' conceptions of physical situations may be context dependent (Vosniadou, 2014). Therefore, it was important to employ the same or similar physical situations in the MAFA as those that were used in prior research on students' misconceptions. The physical situations addressed by the questions were included as a third part of the test domain. The complete test domain was defined by the physical situations employed, the parts of Newton's first and second laws that they addressed, and the misconceptions that they identified. The physical situations served as the stems for knowledge items and guided the structure of the test. Before knowledge items were written, a table was created to show the alignment between each physical situation and the other two parts of the test domain—the domain for knowledge items and the domain for reason items—that could be assessed by each. This table served as the initial draft of the test blueprint. The test blueprint for the MAFA was revised as items were created. The version of the test blueprint corresponding to the initial version of the MAFA is given in Figure 3.3. Physical situations that also served as the basis for FCI questions are noted in the blueprint. The researcher, an experienced physics teacher, created a single knowledge item for each physical situation listed in Figure 3.3. True/false items that targeted each misconception in the table were composed to follow each knowledge item. Some reason items targeted a single misconception and others targeted multiple misconceptions. Between two and four reason items were created for each knowledge item to give a total of 19 knowledge items and 55 reason items in the item pool.

Figure 3.3

Test Blueprint for MAFA

Areas of Domain for Knowledge Items							Physical Situation	Misconceptions for Reason Items					
1a	1b	1c	2a	2b	2c	2d		M1	M2	M3	M4	M5	M6
	X			X			Coin moving upward (FCI 5)	X		X	X		
X	X		X	X			Coin toss at top of path	X		X	X		
		X				X	Rocket turn on engines	X					
X							Rocket turn off engines	X				X	
X	X						Book at rest on table (FCI 12)				X		X
X	X						Book hanging from string				X		X
X	X						Book on hand			X	X		
	X	X		X			Pendulum on upward path	X		X	X		
	X	X		X			Pendulum at bottom of swing	X		X	X		
	X	X	X				Person on rope swing lets go at bottom	X		X			
		X				X	Ball being swung in a circle on a string as seen from above (FCI 4)—before string breaks	X	X		X		
X		X				X	Ball being swung in a circle on a string as seen from above (FCI 4)—after string breaks	X	X	X		X	
X		X					Child on water slide as seen from above	X	X	X		X	
X		X					Puck through curved tube on exit (FCI 10)	X	X	X		X	
	X			X			Ball tossed along parabolic path--upward	X		X	X		
X							Elevator upward at constant speed	X		X	X		
	X			X			Elevator downward and speeding up	X		X			
X							Box being pushed across the floor at constant speed (FCI 28 & 29)	X		X	X		X
	X			X			Box sliding across floor	X		X	X		
							TOTAL	16	4	14	13	4	3

Item Evaluation

Items in the item pool were evaluated in four phases. For all phases, the MAFA was presented to participants using Qualtrics—an online survey platform. At the end of each phase, data from the evaluation were used to revise the items before the next phase of research began. In the first phase, five experienced physics instructors reviewed the items for clarity and for alignment with the misconceptions. The final three phases of the research all involved undergraduate students who had completed no more than two semesters of university level physics courses. In the second phase, four students reviewed the items for clarity. In Phase 3, the pilot test, 100 undergraduate students completed the MAFA. These data were used to choose which items to include in the final version of the MAFA. In the fourth phase, 349 additional students completed the final version of the MAFA. Data from the Phases 3 and 4 were combined for the final analyses. Next, each phase is described in more detail.

Phase One: Expert Review

Participants. In Phase 1, the 19 knowledge items and 55 reason items in the item pool were evaluated by experienced physics instructors for item quality. Five local high school and university instructors who had been teaching introductory physics for at least five years were invited to participate by email. All five physics instructors were acquainted with the researcher and all agreed to review the items. Instructors' teaching experience ranged from 5-15 years with a mean value of 10 years. All instructors had taught at least one introductory level physics course every year they had taught. Three of the instructors taught at local high schools and two taught at Virginia Tech. One of the university instructors was a former high school physics instructor and the other had taught exclusively at the university. Courses taught to high school students included Conceptual Physics, College Prep Physics, and college-level courses such as IB Physics

SL, AP Physics (including levels 1, 2, B, and C), and dual-enrolled algebra-based physics. Courses taught to college students were the first and second semesters of calculus-based Physics for scientists and engineers. All instructors had either a B.A. or a B.S. in Physics. Four of the instructors also had post baccalaureate degrees. Three instructors had a M.Ed. in Curriculum and Instruction and one instructor had a second B.S. in Math as well as a Ph.D. in Physics. The instructors were not compensated for their time.

Data Collection. Phase 1 data were gathered through an online survey platform. Instructors were provided a link to a modified version of the MAFA. They were asked to consider two things as they completed the assessment: the clarity of the items and whether answers to each item might indicate that a student had one of the six misconceptions. For CDAs, a table that shows the attributes needed to correctly answer each item is called a Q-matrix. For the MAFA, the attributes are the six misconceptions, and the Q-matrix indicates which misconceptions are required to answer each reason item in a certain way. The author constructed an initial Q-matrix based on existing research and her own experience as a physics instructor. Instructors were presented with each knowledge item and its associated reason items in the same format as students. However, each set of items was followed by two additional questions—one that asked the instructors to provide feedback about item clarity and another in which they were asked if they agreed with the suggested alignment between misconceptions and item responses in the Q-matrix. If they did not agree with the alignment, the instructors were asked the following question: "Please explain why you chose 'No' and/or provide suggestions to better align [the question] with the misconceptions. This could involve revising the question, choosing different misconceptions, or both." Responses to these questions were used to revise the wording of some items for clarity and to revise the alignment of some item responses with misconceptions in the

Q-matrix. At the end of the assessment, the instructors were asked to provide any suggestions they had for revising the misconceptions.

Phase Two: Think-Alouds

In the second phase of data collection, four undergraduate students who had completed no more than two semesters of college level physics courses performed think-alouds in which they completed the online assessment in the presence of the researcher while voicing their thoughts about each item aloud. Participants were recruited through flyers that were displayed on the Virginia Tech campus. The first four qualified students who contacted the researcher about participating were chosen. Think-alouds are a specific example of cognitive interviewing—a method used to gather validity evidence for assessment items—that has been suggested by researchers to gather validity evidence for response processes that respondents are otherwise assumed to use in responding to questions (Kane 2006; Messick, 1995). Participants met the researcher in a quiet room on campus at a time and location that were convenient to the participant. Each think aloud session lasted between 1.5 and 2 hours and participants were compensated at the rate of $12 per hour by the researcher.

The structure of the think-aloud process was based on a sample protocol used to test U.S. Census surveys which is described in Chapter 7 of Dillman et al. (2014). First, the researcher explained that the respondent was being asked to go through these questions to make sure that people who take the assessment understand the questions. Because of this, they were asked to explain their thinking out loud as they take the test. Next, respondents practiced this technique using a sample question. In addition to familiarizing respondents with the technique, conversation during the practice question was intended to make the respondents more comfortable with the think-aloud experience and move the emphasis of the experience from

answering items correctly to helping to identify problems with the way that the items were written (Dillman et al., 2014). Participants accessed items through a Qualtrics survey on the researcher's computer. The researcher audiotaped and took notes during each session. Data gathered during Phase 2 were analyzed and used to revise the items for greater clarity before proceeding to Phase 3.

Phase Three: Pilot Test

Participant Recruitment and Data Collection. Phase 3 was the pilot test for the assessment. The revised version of the test was piloted with 100 participants. All participants that spent at least eight minutes between opening and submitting the assessment (as indicated by the time stamp in the survey) were compensated $5 for their time. (The researcher was able to read all material and complete the assessment in nine minutes.) Submissions that took less time were considered invalid and were not included in the data set. This phase of data collection began near the end of the Spring 2019 semester. Participants were recruited using multiple methods. First, flyers advertising the study were posted in academic buildings on campus. Second, the researcher emailed information about the study to instructors of introductory biology, chemistry, and physics courses at Virginia Tech and asked the instructors to share the information with their students. Introductory courses were defined as courses in a sequence for which there was not a same-subject prerequisite for the first course in the sequence. For instance, Physics 2215 was considered introductory because the only prerequisite was a mathematics course and Physics 2216, the second course in the sequence, was considered introductory as well. Instructors of introductory science courses were identified using the course timetable for the semester. Interested students emailed the researcher and the researcher replied with an email that described the research (including required IRB information such as who to contact with questions) and

included a digital link to take the assessment. This process was repeated with the modifications described below for courses taught in Summer 2019, Fall 2019, and Spring 2020.

During data collection, the recruitment protocol was modified multiple times due to a low response rate. I received approval for each modification to the research protocol from the appropriate Institutional Review Board (IRB) before implementing each change. Next, the modifications to the recruiting process that occurred during Phase 3 are listed in chronological order along with the date each was implemented. In July 2019, recruitment was expanded by posting flyers at off-campus locations such as grocery stores and coffee shops. In August 2019, emails announcing the study were sent to the leaders of fraternities, sororities, and the corps of cadets at Virginia Tech. In September 2019, I began to conduct in-person recruiting in which I stood in front of the campus library and invited students to participate. In October 2019, I also added instructors of social science courses to the list of instructors to whom information about the study was sent. The required number of 100 participants was reached in January 2020.

Data Analysis. Responses were analyzed using Classical Test Theory (CTT) to identify poorly performing knowledge items and a correlation matrix to measure the extent to which answers to reason items consistently identified misconceptions. Poorly performing knowledge items were those that had very high or very low item difficulties, low or negative discrimination values for the correct answer, and/or distractors with positive discrimination values. Reason items were designed to measure the presence of six misconceptions. One indication that they are doing this well would be that students would answer items that measure the same misconception in similar ways. Correlations between items were calculated to look for high correlations between items that measured the same misconceptions and low correlations between items that measured different misconceptions. Initially, it was thought that the final version of the MAFA

would have 10 knowledge items and their associated reason items to ensure that the time required to take the test was under 30 minutes. However, students completed the test more quickly than had been anticipated. Therefore, I decided that there was no need to remove well-performing items merely to shorten the assessment. Only one knowledge item and five reason items were found to perform poorly. These items were removed from the test to create the final version of the MAFA which was composed of 18 knowledge items and 46 reason items.

Phase Four: Field Test

During Phase 4, the field test, the data needed to model responses to the assessment using Item Response Theory (IRT) for the knowledge items and a diagnostic cognitive model (DCM) for the reason items were gathered. Two questions from the FCI were added to the end of the MAFA to gather validity evidence. These questions were chosen because they used the same physical situations as two questions on the MAFA and because answers to the distractors were aligned to the same misconceptions that were assessed by the MAFA items. As in Phase 3, participants were offered $5 to compensate them for their time. It was projected that a total of 400 additional responses would be required to estimate the final test parameters. Despite extended recruitment efforts, participant response rate remained relatively low. Therefore, in Spring 2020 the researcher requested and received permission to recruit participants from six additional public universities in Virginia: Christopher Newport University, George Mason University, James Madison University, Radford University, University of Virginia, and University of Virginia-Wise. Data for Phase 4 were collected during Spring 2020, Summer 2020, and Fall 2020. The same recruitment protocols that were used in Phase 3 were used in Phase 4 with the exception of in-person recruiting due to the COVID-19 pandemic which caused many campuses to send students home and made person-to-person contact unwise. In total, 1692 initial

emails were sent to contact instructors during these three academic sessions. However, the response rate remained low. The possibility of combining data from Phases 3 and 4 for the final analysis was explored. Because there were only minor changes between the two versions of the test, it seemed unlikely that they would have a significant effect on students' responses. Therefore, it was decided to combine the data from Phases 3 and 4 for the final analysis. Finally, these data were used to estimate item and person parameters for each set of items and to compare answers to the MAFA and the FCI questions. Because model fit was poor when all items were included in each model, items were eliminated from the final IRT model and DCM to improve model fit.

Results

Phase 1: Expert Review

In Phase 1 of the research, five experienced physics instructors provided feedback on the content and composition of the items in the item pool and on the extent to which responses to the reason items would indicate the presence of one or more of the six misconceptions as shown in the Q-matrix. Instructors' suggestions regarding item clarity were used to revise the items and reduce the probability that the questions, answers, and diagrams would be misinterpreted. Suggestions about the alignment of misconceptions and item responses revealed possible assumptions about some items that I had not considered, but that I agreed with. These were as follows:

1. In the original version of a question about a coin tossed into the air, one of the reason items read: "The net force decreases as the coin gets higher" and it was suggested that answering this question with "True" aligned with misconceptions M1 and M3. Three

instructors pointed out that students who know about the Law of Universal Gravitation would know that the gravitational forces do decrease as objects get farther apart and might not assume that this change is negligible for the coin. Therefore, the question was rewritten to read: "The coin slows down because the net force decreases as the coin moves upward" and the alignment was left the same.

2. In a question that referred to a rocket changing direction because it pushes fuel into space, "fuel" was changed to "fuel exhaust".
3. Multiple items asked "which force(s)" acted on an object. It was suggested that this wording implied that multiple answers could be chosen, so the wording was changed to "what forces". Also, nouns were substituted for pronouns in two items.
4. In multiple items, a statement was added to assume no air resistance or that it was negligible.

Data from Phase 1 were also used to revise the Q-matrix. The entry for one item was changed to indicate the presence of an additional misconception that was pointed out by one of the instructors. Instructors disagreed with other entries in the Q-matrix, but they made suggestions to revise the wording of the items and misconceptions rather than the Q-matrix to bring the items and misconceptions into alignment. This resulted in additional revision of the reason items and the misconceptions.

Phase 2: Think-Alouds

During Phase 2, four students performed think-alouds while completing the MAFA. The researcher audiotaped the sessions and took notes. The think-alouds served two purposes. First, students identified wording that they found confusing or thought might be potentially confusing

to others. For instance, two participants pointed out that not all students would understand the term “net force” and suggested that it be replaced with “total force”. Second, participants explained their reasons for choosing and eliminating answers to each question. The researcher found that students were interpreting the questions as predicted. This provided evidence for content validity.

Phase 3: Pilot Test

Data collected during Phase 3 were analyzed to identify poorly performing items so that they could be deleted from the final version of the MAFA. Poor item performance could include knowledge items which almost everyone or almost no one answered correctly, distractors which were chosen by few respondents, knowledge items that were more likely to be answered correctly by respondents with lower overall scores, and reason items associated with the same misconceptions that were only weakly correlated. Because the MAFA knowledge items and reason items work together, decisions about whether to retain or delete each item were made only after considering the performance of all items in the set. Only complete submissions for which submissions which were made at least 8 minutes after opening the survey were included in the analysis. It was estimated that this was the minimum time required to read and respond to all the items. Responses were analyzed using jMetrik software (Meyer, 2018).

The knowledge items were analyzed using the Classical True Score Model (CTSM). Although the final analysis of items was not done under the CTSM, it was useful to provide simple measures of item difficulty and discrimination using the pilot test data to ensure each item included on the test provided useful information. The researcher considered dropping items with difficulties near 0.0 or 1.0 as well as those with discriminations less than 0.3. There were four items that had these characteristics. Each item was inspected for unforeseen problems and to see

if the reason questions that followed it added valuable information to the misconceptions profile to decide whether to drop or retain the item. The statistics and decisions for each item are described in Table 3.2.

The performance of the distractors for each knowledge item was also analyzed. Two values, the proportion of respondents who chose the distractor (p) and a discrimination index (r_{pbis}), were calculated for each distractor. Distractors with very low values for p (chosen by few respondents) may provide little information for the total test score and those with very high values for p may be a second correct answer. This process identified five knowledge items for which one of the distractors was chosen by no one. Although the answer choices could have been dropped without losing information, it was decided to leave them so that all knowledge items would have four answer choices. On a traditional test, discrimination values should be negative for most distractors as this would indicate that students who have a high test score are less likely to choose the distractor on the item. Distractors with positive discriminations were examined for potential problems. Decisions about whether to drop items were made only after considering the performance of both the knowledge item and the reason items associated with it because the reason items associated with low discrimination distractors might still provide valuable information about misconceptions. Table 3.2 lists the statistics and decisions that were made for potentially problematic items.

The final step in analyzing Phase 3 data was to calculate correlation coefficients for each reason item. It was expected that reason items which measured the same misconception would be highly correlated. The possibility that reason items that measured different

Table 3.2

Items Considered for Removal During Phase 3

Item ID	Item Description	Problem	Decision and Reason
		Knowledge Items	
Knowledge	Rocket in space--engines turn on	Low discrimination (r = -0.0114)	Left in test—required precursor to next item
Knowledge	Ball on string before string breaks	Low discrimination (r = 0.2151)	Left in test—Associated reason items are strong.
Knowledge	Spiral water slide	Low discrimination (r = 0.1068)	Dropped—Some confusion was also noted during Phase 2. Associated reason items also dropped.
Knowledge	Elevator moving upward at constant speed	Low discrimination (r = 0.2123)	Left in test—Associated reason items are strong.
		Reason Items	
Reason	Only force on ball after string breaks is gravity.	Low correlation with other items measuring same misconception	Left in test, but alignment of certain responses with Q-matrix deleted—Interpretation of alignment was only problematic for some responses to associated knowledge item.
Reason	No gravity acts on book at rest on table because it is at rest.	Very low or negative correlations with other items measuring same misconception	Dropped—May be that students interpreted this situation differently because book was supported by table rather than string or hand.
Reason	No gravity acts on pendulum at lowest point in swing AND No gravity acts on coin as it moves upward	Negative correlation	Left in test—Both items correlated well with other items that measured the same misconceptions.
Reason	Constant upward force on elevator moving upward at constant speed	Low correlation with all other items measuring same misconception	Dropped—Item was wordy and may have been misinterpreted.
Reason	Floor of elevator exerts an upward force on person's feet as elevator moves downward.	Negative correlation with other items measuring same misconception.	Dropped—Item was wordy and may have been misinterpreted.

misconceptions might also be high correlated because some of the misconceptions were closely related was also recognized. Therefore, items that measured the same sets of misconceptions were placed into groups and correlation coefficients were calculated between all item pairs in the group. For instance, all items that measured only misconception one were placed into a single group and all items that measured only misconceptions one and four were placed into a different group. For items aligned with more than one misconception, it was thought that both misconceptions were required to choose these items. Therefore, items that measured only misconception one or misconception four were not placed with items that measured both misconceptions. Reason items that had low or negative correlations with other items in their group are listed in Table 3.2 along with the decisions that were made about each.

Phase 4: Field Test

Participants and Data

Data from Phase 4 (N = 352) were combined with data from Phase 3 (N = 97) for the final analysis and model fitting. The "loss" of three cases from the Phase 3 data was due to the researcher's decision to increase the minimum time for completion from 8 minutes to 9 minutes. Nine minutes was the time needed for the researcher to read each question without taking time to think about what the correct answer might be. All respondents that had no correct answers to knowledge items and most that had only one or two correct answers fell into this group. The highest performance that was eliminated had five correct knowledge items. Because the time stamp on responses recorded only start and end time and the assessment link allowed respondents to leave and return to the assessment over a period of two weeks, comparing mean times for completion was not meaningful. There was some concern that these data might differ because the Phase 4 data collection had been extended to additional schools. Therefore, the

Phase 4 data were compared to the Phase 3 data in terms of participant demographics and number of knowledge items answered correctly. The results of this analysis are shown in Table 3.3.

Table 3.3

Comparison of Participants by Phase of Research

Question	Response	Phase 3 Data (N = 97) Number / Percent	Phase 4 Data (N = 352) Number / Percent	All Data (N = 449) Number / Percent
School Year	Freshman	44 / 45.4%	167 / 47.4%	211 / 47.0%
	Sophomore	21 / 21.6%	102 / 29.0%	123 / 27.4%
	Junior	15 / 15.5%	49 / 13.9%	64 / 14.3%
	Senior	15 / 15.5%	34 / 9.7%	49 / 10.9%
	Prefer not to answer	2 / 2.1%	0 / 0%	2 / 0.4%
	Total	97 / 100.0%	352 / 100.0%	449 / 100.0%
Sex	Female	47 / 48.5%	235 / 66.8%	282 / 62.8%
	Male	49 / 50.5%	117 / 33.2%	166 / 37.0%
	Prefer not to answer	1 / 1.0%	0 / 0.0%	1 / 0.2%
	Total	97 / 100.0%	352 / 100.0%	449 / 100.0%
Physics courses completed in high school	None	22 / 22.7%	108 / 30.7%	130 / 29.0%
	Regular/Honors/Conceptual only	36 / 37.1%	158 / 44.9%	194 / 43.2%
	At least one AP or IB course	39 / 40.2%	86 / 24.4%	125 / 27.8%
	Total	97 / 100.0%	352 / 100.0%	449 / 100.0%
Number of semesters of college level physics completed	None	79 / 81.4%	304 / 86.4%	383 / 85.3%
	One	9 / 9.3%	27 / 7.7%	36 / 8.0%
	Two	9 / 9.3%	21 / 6.0%	30 / 6.7%
	Total	97 / 100.0%	352 / 100.0%	449 / 100.0%
Number of Knowledge Items Correct	Minimum	4	2	2
	Maximum	18	18	18
	Mean (SD)	12.15 (4.04)	9.46 (3.60)	10.04 (3.85)

The participant demographics are similar in terms of school year and number of semesters of college physics courses completed. They differ in two things. First, about 67% of the Phase 4 participants are female while only 49% of Phase 3 participants are female. The other difference is in the physics courses that participants completed in high school. Phase 4 participants completed fewer high school physics courses-especially AP or IB physics courses. A third difference between groups is the mean number of knowledge items answered correctly. Participants in the Phase 4 group had a lower mean score (9.46/18.00) than those in the Phase 3 group (12.15/18.00). Despite these differences, all the students came from the target population for the test--undergraduate students who have completed no more than 2 semesters of college level physics. The combined sample was more representative of the population for whom the test is intended. It was anticipated that the combined data set would result in more precise parameter estimates.

Model Fit for Knowledge Items

Knowledge items were modeled using item response theory (IRT). The first steps were to test the model assumptions of unidimensionality and local item independence which were done using IRTPro software (Cai et al., 2017a). The next steps--comparing the fit of a 1-PL model to a 2-PL model and using the best model to estimate item parameters and person ability estimates--were done using jMetrik software (Meyer, 2018). This section describes the results of each of these processes.

Testing Model Assumptions. The IRT models used to analyze responses to the MAFA knowledge items (1-PL and 2-PL) are built upon the assumption that the test is unidimensional—that responses to all items are explained by the same latent construct. The MAFA was hypothesized to measure a single latent construct: knowledge of Newton's first and second laws.

Exploratory Factor Analysis (EFA) was used to compare model fit for a one-dimensional model and a two-dimensional model. This type of factor analysis is used when there is no a priori hypothesis about which items may load onto which factors. The EFA process identifies which items load onto the same latent construct which allows the researcher to compare the fit of models in which different numbers of latent constructs are specified. The models specified quartimax rotation which allows multiple factors to be partially correlated while producing a small number of factors.

The model fit for the one- and two-dimensional models were compared using the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwartz, 1978). The measures for the one-dimensional model (AIC = 9634.29 and BIC = 9782.15) were higher than for the two-dimensional model (AIC = 9480.69 and BIC = 9698.36). Lower values generally indicate a better fit. However, it is important to consider more than fit indices when deciding on the dimensionality of a test. For instance, factor loadings should make sense in terms of question content rather than shared item stems, similar difficulty levels, or other less salient constructs. Next, possible explanations for the fit of the two-factor model were considered by looking at factor loadings, local item dependence, and item characteristic curves. Chen and Thissen's (1997) local dependence chi-square statistic was used to assess local item dependence with values above 10.0 considered indications of dependence as recommended by the authors. Finally, item characteristic curves were inspected to ensure that they were monotonically increasing showing that students with higher abilities were more likely to answer each item correctly.

Table 3.4 shows the factor loadings, proportion correct, and topic for MAFA knowledge items in the two-factor model. A factor loading of 0.30 or higher was used to determine if an item loaded onto a given factor (Scott & Schumayer, 2012). Six items loaded onto factor 1,

Table 3.4

Quartimax Rotated Loadings for Two-Factor EFA

Item	Factor 1		Factor 2		Proportion Correct	Topic
	λ_1	*SE*	λ_2	*SE*		
Q7	**.96**	**.19**	**-.07**	**.40**	**.80**	Which forces on a book at rest on hand
Q18	.91	.18	-.01	.36	.33	Number of forces on box pushed at constant speed
Q5	.88	.20	.10	.35	.62	Which forces on book at rest on table
Q6	**.78**	**.17**	**.00**	**.32**	**.69**	Which forces on book hanging on string
Q12	.45	.19	.34	.29	.43	Path of ball on string swung in circle and released
Q14	.36	.22	.15	.17	.73	Path of puck leaving end of curved tube
Q1	**.03**	**.25**	**.81**	**.03**	**.40**	Which forces on a tossed coin going up
Q15	**-.08**	**.28**	**.73**	**.16**	**.39**	Number of forces on juggled ball going up
Q2	.17	.27	.61	.20	.48	Which forces on tossed coin at top of path
Q8	-.01	.23	.54	.16	.56	Number of forces on pendulum going up
Q11	**-.27**	**.18**	**.51**	**.14**	**.49**	Number of forces on ball on string swung in circle
Q9	.25	.24	.43	.23	.58	Number of forces on pendulum at bottom of swing
Q16	.07	.21	.40	.16	.65	Which forces on elevator going up
Q17	.22	.21	.35	.20	.43	Which forces on person in elevator
Q19	.19	.18	.29	.18	.47	Number of forces on sliding box
Q10	.26	.16	.11	.18	.65	Path of person on rope swing
Q4	.29	.18	.10	.20	.64	Path of rocket in space after engine turned off
Q3	**-.13**	**.15**	**.11**	**.16**	**.69**	Path of rocket in space after engine turned on

Note. Items are ordered by strength of loading on each factor. Items in bold were eliminated from final scoring.

eight items loaded onto factor 2, and four items did not load onto either factor. Three of the items that loaded onto factor one—Q5, Q6, and Q7—asked about the forces that act on a book at rest. Two of the items, Q5 and Q7, had a local dependence chi-square statistic greater than 10.0. The three items all asked about forces supporting a book and this similarity may have accounted for the correlation and possible local item dependence between the three items. The three items aligned to the same two areas of the test domain for knowledge items, 1a and 1b, but differed in difficulty. The two easier items, Q6 and Q7, were eliminated from the IRT models. The remaining items in factor one were not logically related to either each other or the book questions. Similarly, three questions that loaded onto factor two—Q1, Q2, and Q15—are about the forces that act on an object that has been tossed into the air. In this case, all pairs of the three items had local dependence chi-square statistics greater than 10.0. The only difference between the three questions was the object being tossed, whether the object was tossed straight upward or at an angle, and where along the path the object was located. It was decided to eliminate two of the questions—Q1 and Q15--from the IRT models. These two questions aligned to only parts 1b and 2b of the test domain for knowledge items while Q2 aligned with parts 1a, 1b, 2a, and 2b. The other questions that loaded onto factor two were not logically related to the first three questions and were left in the IRT models. Two additional items, Q3 and Q11, were removed after inspecting the item characteristics curves. The curve for Q3 was monotonically decreasing indicating that respondents with overall higher scores were less likely to answer the item correctly. This was not surprising as the Q3 (about a rocket in space) had a discrimination near zero in the pilot test but was left in the test because it was the precursor to the following question. The curve for Q11 had a very low slope which meant that it would do little to

differentiate respondent abilities. Based on these results, it was decided to remove these six items from the data and estimate the two EFA models again for the remaining twelve items.

Of course, no test will be perfectly unidimensional, but the factors must make sense. According to Kane (2006), "the interpretation depends on a combination of formal mathematical modeling and subjective judgements that tie the model to observable phenomena" (p. 41). EFA conducted for the remaining 12 items on the MAFA showed mixed results when comparing the one- and two-dimensional models. The AIC was slightly lower for the two-dimensional model (6607 compared to 6640) while the BIC was slightly lower for the one-dimensional model (6739 compared to 6751). Factor loadings for the one-dimensional model were greater than 0.30 for all twelve items. Factor loadings for the two-dimensional model showed four items with loadings greater than 0.30 for factor one, seven items with loading greater than 0.30 for factor two, and one item for which the greatest loading was 0.25. However, few of the items that loaded onto each factor showed a logical relationship to each other. Therefore, it was decided to continue the analysis with 12 items and a unidimensional IRT model.

Comparing IRT Models. Two IRT models—a 1-PL model and a 2-PL model—were fit to the knowledge item responses using the jMetrik software package (Meyer, 2018). A prior lognormal distribution with mean of 0 and standard deviation of 0.5 was applied to discrimination parameters and a prior beta distribution with mean of 0 and standard deviation of 1.0—the default for the software—was applied to item difficulty parameters. The two models were compared at the item level and for overall model fit using multiple measures. Based on both considerations, the 2-PL model was deemed a better fitting model for the MAFA knowledge items. Next, I compare the item level fit for the two models. This is followed by a comparison of overall model fit for the two models.

At the item level, a chi-square statistic is provided that compares the overall distribution of responses to an item compared to the expected distribution for each sum score. Chi-square item level statistic *p*-values less than .05 indicate negligible misfit and *p*-values less than .01 indicate more serious misfit (Cai et al., 2017b). Comparing the 1-PL and 2-PL models, for the 1-PL model there were three items with chi-square values less than .05—two of these less than .01—and for the 2-PL model there were no items with chi square values less than .05. For overall model fit, AIC and BIC were used with smaller values indicating better fit. The AIC was smaller for the 2-PL model (3345 for the 2-PL versus 3370 for the 1-PL) and the BIC was smaller for the 1-PL model (3420 for the 1-PL versus 3444 for the 2-PL). The 2-PL model was deemed a better fit overall. Item parameter estimates and chi-square item level statistics for the 2-PL model are given in Table 3.5. Estimates for item difficulty range between -1.29 (Q14) and 0.48 (Q18). Estimates for item discrimination parameters vary from 0.58 (Q10) to 2.63 (Q5). Standard errors for estimates are almost all less than or equal to 0.20 and no chi-square item level fit statistics are significant at the $p < .05$ level.

Person scores were estimated using the expected a posteriori (EAP) method. In EAP estimation, each response pattern is assigned an ability estimate. A twelve-item test such as the MAFA has $2^{12} = 4096$ possible response patterns. However, only 345 different response patterns were seen in the sample. Given the sample size of 449, this means that few response patterns occurred more than once. The most common response patterns were all items correct (n = 31), all items correct except for Q10 (n = 7), and all items correct except for Q17 (n = 7). All other response patterns that were present in the data set appeared between one and three times. Figure 3.4 shows the distribution of person scores for the 449 respondents. Ability estimates ranged from -1.97 to

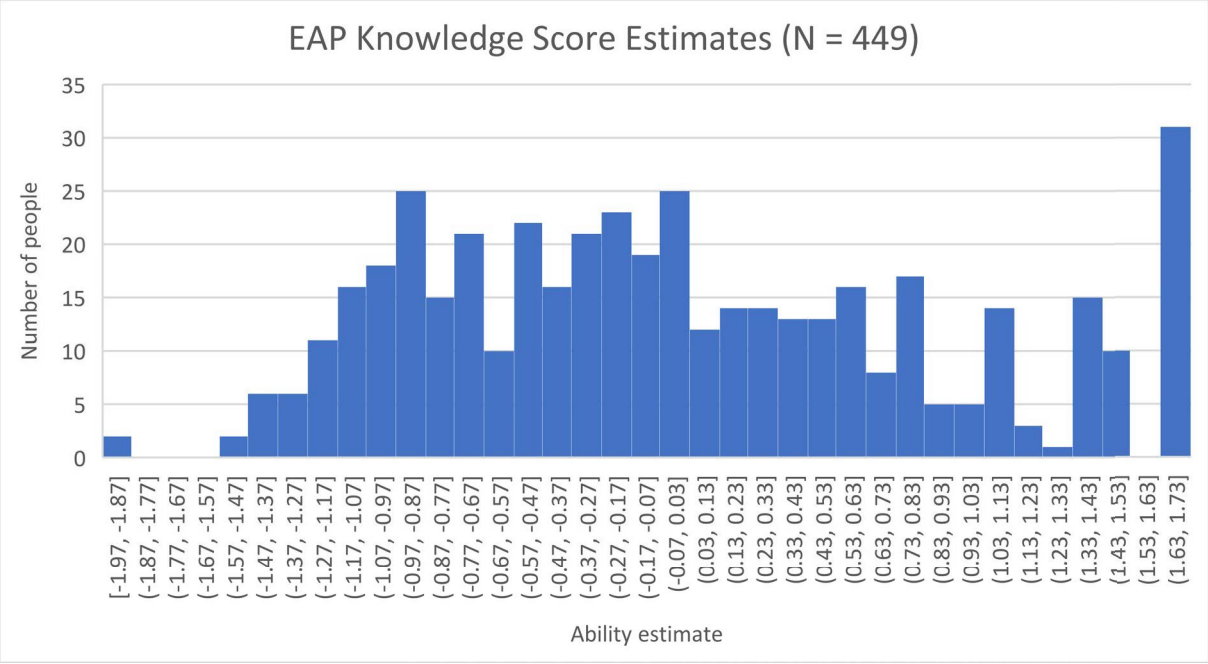
Table 3.5

Item Parameters and Fit Statistics for 12-Item 2-PL Model

Item	Item Difficulty		Item Discrimination		Item level X^2		
	β	*SE*	α	*SE*	X^2	*df*	*p*
Q2	.05	.09	1.17	.13	7.17	8	.5183
Q4	-.91	.18	.74	.11	12.38	8	.1351
Q5	-.40	.05	2.63	.26	3.44	8	.9037
Q8	-.44	.18	.59	.10	6.03	8	.6438
Q9	-.42	.11	1.03	.13	5.97	8	.6506
Q10	-1.17	.25	.58	.10	13.39	8	.0990
Q12	.26	.08	1.46	.15	8.15	8	.4190
Q14	-1.29	.19	.88	.13	6.02	8	.6454
Q16	-1.06	.22	.63	.11	6.98	8	.5384
Q17	.32	.12	.92	.12	7.05	8	.5316
Q18	.51	.06	2.31	.22	13.69	8	.0902
Q19	.16	.14	.72	.10	6.93	8	.5445

Figure 3.4

EAP Score Estimates for 12 Knowledge Items



+1.70 with a mean value of 0.00. Standard errors ranged between 0.45 and 0.66 with a mean value of 0.51.

For IRT models, a plot of the test information function, the test information curve, shows where ability estimates have the greatest and least precision. The test information curve for the twelve knowledge items included in the final 2-PL model is given in Figure 3.5. The greatest values for information are for ability estimates between approximately -1.0 and 1.0 with maximum information given at -.20. Measurements are most precise for average students.

Model Fit for Reason Items

The MAFA consists of sets of related knowledge and reason items. Responses to the knowledge items were fit to a 2-PL IRT model which provided a single measure of knowledge about forces in relation to Newton's first and second laws. Reason items were designed to elicit information about the presence or absence of six specific misconceptions about Newton's first and second laws. Therefore, they were fit to a diagnostic cognitive model, the DINA model, designed to provide a profile of a set of constructs or skills. For the MAFA, the constructs or skills that were measured were the six misconceptions. Each reason item was aligned with between zero and two misconceptions. The MAFA is shown in Appendix A and the alignments are shown in the Q-matrix which is given in Appendix B.

The scoring of reason items is inherently backwards. Because the items were designed to identify students who possessed misconceptions, "correct" answers did not indicate correct knowledge. They indicated the presence of misconceptions. The scoring for reason items (see Table 3.6) was based on the q matrix that was specified in Phase 1. For all items, a score of "1" indicates the presence of all misconceptions associated with the item and a score of "2" indicates

the absence of at least one of the misconceptions associated with the item. For most reason items, a response of "True" indicates the presence of misconceptions. For some items, however, a response of "False" indicates the presence of misconceptions. Finally, there were some items for

Figure 3.5

Test Information Curve for 12 Knowledge Items

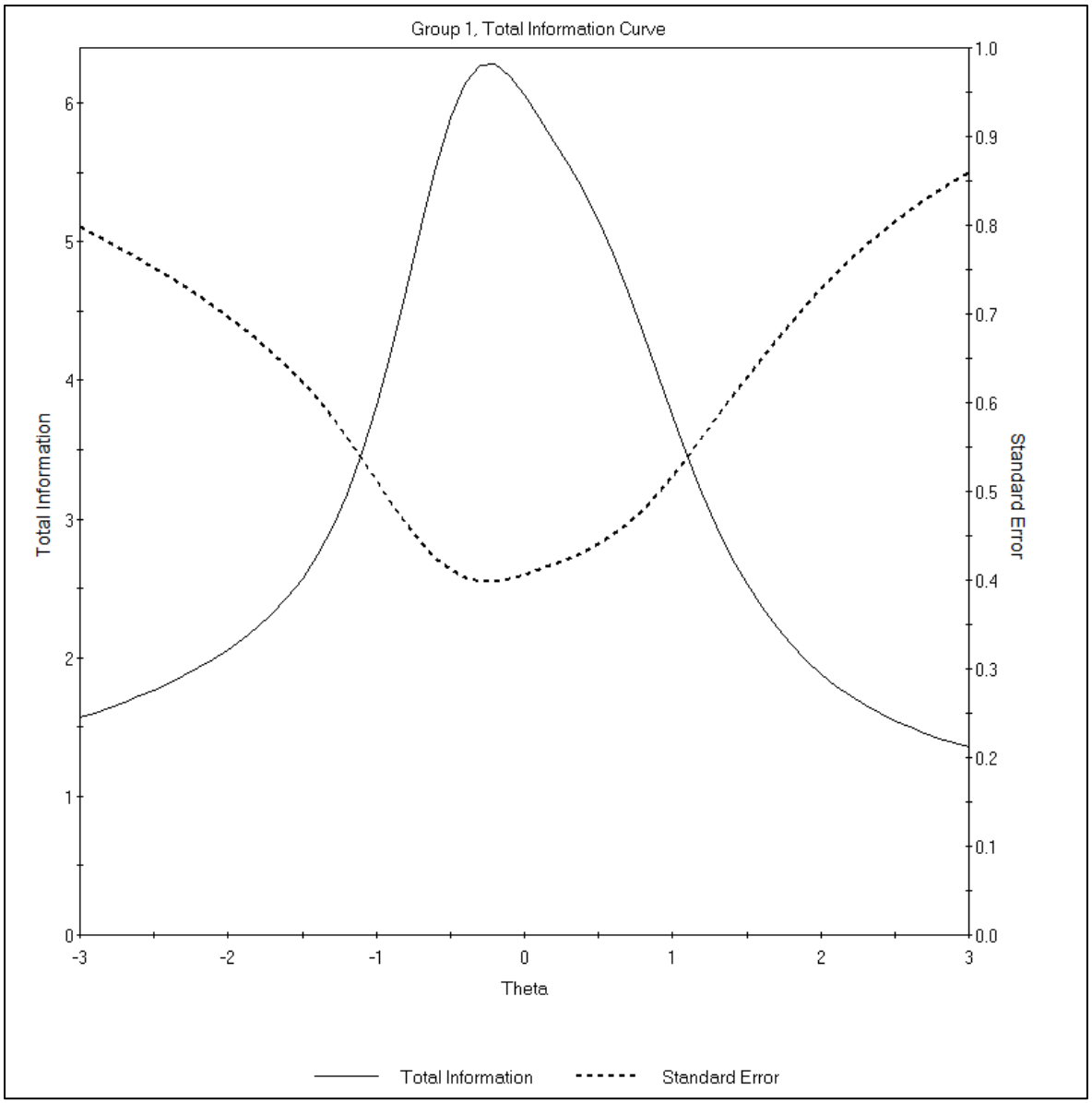


Table 3.6

Scoring Instructions for Reason Items

Knowledge Item	Reason Item	Knowledge Item Response	Score[a] for "True"	Score[a] for "False"
Q1	Q1.3	A	2	2
		B, C, or D	1	2
	Q1.4	A or C	2	2
		B or D	1	1
Q3	Q3.1	A, B, C, or D	2	1
Q4	Q4.1	A, C, or D	2	2
		B	1	2
Q6	Q6.3	A, B, C, or D	2	1
Q7	Q7.3	A, B, C, or D	2	1
Q8	Q8.3	A	2	2
		B, C, or D	1	2
Q12	Q12.2	A or C	2	1
		B or D	2	2
Q14	Q14.2 [b]	A or C	2	2
		B or D	1	2
	Q14.2a[b]	A or C	1	2
		B or D	2	2
Q15	Q15.2	A	2	2
		B, C, or D	1	2
	Q15.3	A	2	2
		B, C, or D	1	2
Q18	Q18.2	A, B, C, or D	2	1
Q19	Q19.2	A, B, C, or D	2	1

Note. The following reason items are scored as T = 1 and F = 2 for all responses to the associated knowledge items: Q1.1, Q1.2, Q2.1, Q2.2, Q3.2, Q4.2, Q5.2, Q5.3, Q6.1, Q6.2, Q7.1, Q7.2, Q8.1, Q8.2, Q9.1, Q9.2, Q9.3, Q10.1, Q10.2, Q10.3, Q11.1, Q11.2, Q11.3, Q12.3, Q14.3, Q14.4, Q15.1, Q16.1, Q17.1, Q18.1, Q18.3, Q19.1, and Q19.3.

[a]A score of "1" indicates the presence of all misconceptions aligned with the item and a score of "2" indicates that the absence of at least one of the misconceptions aligned with the item.

[b]Responses to Q14.2 correspond to different sets of misconceptions depending on responses to the associated reason items. The question occupies two rows in the Q-matrix. The rows are labeled Q14.2 and Q14.2a.

which the presence of misconceptions was indicated by a combination of responses to the knowledge item and the reason item together. The scoring for these items was more complicated. For instance, an answer of "A" to knowledge item 20 item indicated the presence of no misconceptions regardless of the response to the associated reason item 24.

Values for the parameter estimates were computed using the R-package CDM (Robitzsch et al., 2020a) and choosing the DINA model. For the DINA model, smaller slipping and guessing parameters indicate better fit (Rupp et al., 2010). In addition to low slipping and guessing parameter estimates, indices for judging item fit include item level RMSEA and item discrimination. Rupp et al. (2010) define item discrimination for the DINA model as the probability of neither guessing nor slipping with higher values indicating stronger items. Generally, RMSEA values greater than 0.10 indicate poor model fit, values between 0.10 and 0.05 indicate moderate fit and values less than 0.05 indicate good fit. The RMSEA values, item discrimination indices (IDI) and slipping and guessing parameters were used to decide whether to keep items in the model. When all 38 reason items were included in the model, the mean RMSEA was 0.104, some items had RMSEA values greater than 0.30, two items had IDI values less than 0.10, and two items had slipping or guessing parameters greater than 0.9. In order to improve model fit, items were deleted from the model in a stepwise fashion starting by removing items that had RMSEA values greater than 0.15, IDI values less than 0.10, or slipping or guessing parameters greater than 0.9 and rerunning the model. This process was repeated for RMSEA values greater than 0.14, greater than 0.13, greater than 0.12, greater than 0.11, greater than 0.105, and greater than 0.10. There was no need to remove items due to large slipping and guessing parameters after the first step because no values exceeded 0.9 after that point. The same

was true for items with IDI less than 0.10. At each step, overall model fit statistics, item parameters, and other model estimates such as correlations of misconceptions were compared.

Many models were compared with different numbers (from 38 to 22) and combinations of items. The model that was judged to have the best overall fit statistics included 27 items. Item parameters, discrimination indices, and RMSEA for the 27 items are given in Table 3.7. Estimates for guessing parameters ranged from 0.000 to 0.429 with a mean value of 0.078. Estimates for slipping parameters ranged from 0.114 to 0.783 with a mean value of 0.444. Item discrimination index values ranged from 0.303 to 0.821 with a mean value of 0.592. The mean RMSEA was 0.076 which is considered a moderate fit.

The R package CDM provides multiple statistics that can be used to judge overall model fit and local item independence. The statistics are based on comparing the expected and observed responses for pairs of items (Robitzsch et al., 2020b). These statistics include: the mean absolute deviation between observed and model-predicted correlations of item pairs (MADcor), the standardized root mean square root of squared residuals (SRMSR) (Maydeu-Olivares, 2013), the mean of absolute deviations of residual covariances times 100 (100*MADRESIDCOV) (McDonald & Mok, 1995), the mean of absolute values of the Q3 statistic (MADQ3) (Yen, 1984), and the mean of absolute values of the centered Q3 statistic (MADaQ3). For each of these statistics, the closer the value is to zero, the better the model fit (Robitzsch et al., 2020b). Values for each of these fit statistics for the 27 reason items were as follows: MADcor = .048, SRMSR = .068, 100*MADRESIDCOV = .683, MADQ3 = .058, and MADaQ3 = .058. Recommendations for cut-off values for determining overall model fit differ. For instance, Maydeu-Olivares (2013) suggests that good model fit is indicated by SRMSR values less than

Table 3.7

Item Parameters and Fit Indices for Reason Items

Item	Guessing Parameter Estimate	Slipping parameter estimate	Item Discrimination Index	RMSEA
Q1.2	0.000	0.783	0.217	0.065
Q1.4	0.096	0.261	0.643	0.115
Q2.1	0.054	0.693	0.253	0.119
Q3.1	0.251	0.229	0.520	0.079
Q4.2	0.429	0.280	0.291	0.109
Q6.2	0.000	0.615	0.385	0.068
Q7.2	0.002	0.671	0.327	0.052
Q7.3	0.067	0.524	0.409	0.066
Q8.1	0.004	0.708	0.287	0.061
Q8.3	0.190	0.177	0.633	0.082
Q9.2	0.000	0.511	0.489	0.068
Q9.3	0.094	0.205	0.701	0.058
Q10.1	0.094	0.307	0.599	0.059
Q10.2	0.036	0.309	0.655	0.063
Q11.1	0.268	0.114	0.618	0.075
Q12.2	0.022	0.482	0.496	0.100
Q12.3	0.084	0.056	0.860	0.008
Q14.4	0.039	0.114	0.847	0.049
Q15.1	0.022	0.699	0.280	0.063
Q15.2	0.031	0.329	0.640	0.108
Q15.3	0.151	0.150	0.699	0.120
Q16.1	0.023	0.675	0.302	0.068
Q17.1	0.129	0.528	0.343	0.095
Q18.1	0.011	0.673	0.316	0.049
Q18.2	0.000	0.674	0.326	0.094
Q19.2	0.009	0.677	0.314	0.080
Q19.3	0.011	0.540	0.449	0.086
Mean	0.078	0.444	0.478	0.076

0.05, but Hu and Bentler (1999) suggests that values up to .08 indicate good fit. The model fit for the MAFA is considered good by the second measure, but not by the first. However, this 27-item model had the lowest overall values for the fit measures of all models that were compared.

In addition to the statistics provided above, the CDM package (Robitzsch et al., 2020a) provides two hypothesis tests, each accompanied by a *p*-value, for overall model fit. The first test, max(X2), is based on chi-square tests of the frequency of expected and observed responses between each set of item pairs. The statistic max(X2) is defined as the maximum of all the chi-square statistics for item pairs. The *p*-value for this statistic is determined using the Holm procedure. The value of this statistic for the MAFA indicated poor model fit (max(X2) = 26.82, $p < .001$). A second statistic, abs(fcor) is the "absolute value of the deviations of Fisher transformed correlations as used in Chen et al. 2013" (Robitzsch et al. 2020b, p.167). The value of abs(fcor) for the MAFA also indicated poor model fit (abs(fcor) = 0.48, $p < .000$). Overall, the 2-PL model showed good fit to the MAFA responses and the DINA model showed moderate to poor fit.

Tetrachoric correlations between the six misconceptions varied widely from a low of 0.20 to a high of 0.97. Values for all pairs are given in Table 3.8. The table shows that there are high correlations between M1 and all other misconceptions. It should also be noted that there are more items that measure M1 than any other misconception. A potential problem with the selection of items in the final model is that there is only one item that measures M2. Different combinations of items were tested to find a set of items that performed well and included multiple items to measure M2, but none were found.

The marginal skills distribution for the MAFA is shown in the left side of Figure 3.6. The model indicates that the majority of students possessed multiple misconceptions. This is also shown by the probabilities for the misconception profiles which are plotted in right side of the figure and which indicate that over 30% of respondents are likely to possess misconceptions M1,

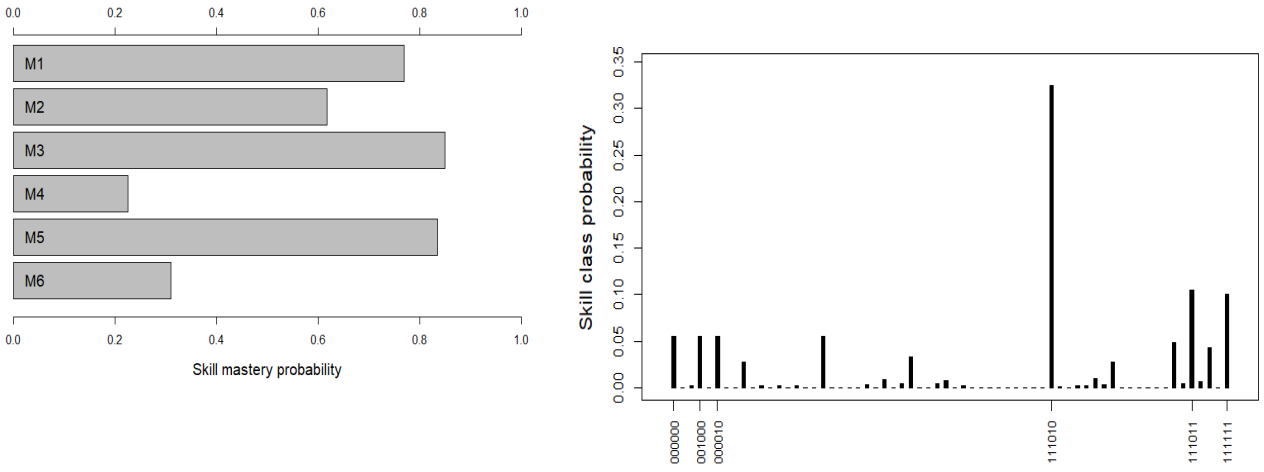
Table 3.8

Correlations Between Misconceptions and Number of Items Measuring Each

Variable (# of items)	1.	2.	3.	4.	5.
1. Misconception 1 (14 items)					
2. Misconception 2 (1 item)	.92				
3. Misconception 3 (6 items)	.80	.68			
4. Misconception 4 (11 items)	.97	.20	.33		
5. Misconception 5 (2 items)	.75	.79	.53	.30	
6. Misconception 6 (4 items)	.83	.23	.28	.73	.45

Figure 3.6

Probabilities of Possessing Misconceptions



M2, M3, M5, and M5; over 10% are likely to possess misconceptions M1, M2, M3, M5, and M6; and over 10% are likely to possess all six misconceptions.

Comparison of Responses to FCI Items and MAFA Reason Items

An additional measure of validity evidence for the MAFA reasons items was provided by comparing responses to two FCI items (included as Q20 and Q21 in the MAFA) and four MAFA items (knowledge item Q1 and reason items Q1.3, Q1.4, and Q18.3) that measured the same misconceptions. The items are shown side-by-side in Figure 3.7. The first FCI item (Q21) was about an object that had been tossed into the air and asked about the forces acting on the object as it moved upward to the top of its path and downward toward the ground. This FCI item aligned well with the first set of questions on the MAFA about an object that has been tossed into the air (Q1, Q1.3, and Q1.4). One difference is that the MAFA item asks only about the first half of the trip as the coin is moving upward. The second FCI item (Q20) was about a box that is being pushed across the floor at a constant velocity. Two responses to this item (*d* and e) aligned with a MAFA knowledge item (Q18.3) that is written around the same scenario.

Table 3.9 gives a comparison of responses to the paired items that would be expected to be answered similarly. For instance, a person who chooses answer *a* ("A downward force of gravity along with a steadily decreasing upward force") for Q21 should also choose answer *d* ("Both an upward force and a downward force") for the Q1. However, not all respondents who choose "Both and upward force and a downward force" for the MAFA item would necessarily be expected to choose the matching response to the FCI item because they might not think that the upward force is steadily decreasing. Therefore, a comparison was made between respondents who chose answer *d* to Q1 and those that chose answers *a*, *b*, or *c* to Q1 for the 36 respondents who chose answer *a* to Q21. A chi-square goodness of fit test was used to determine whether the

Figure 3.7

FCI and MAFA Items Used in Validity Argument

FCI	MAFA
Q21. A boy throws a steel ball straight up. Consider the motion of the ball only after it has left the boy's hand but before it touches the ground, and assume that forces exerted by the air are negligible. For these conditions, the force(s) acting on the ball is (are) a. A downward force of gravity along with a steadily decreasing upward force b. A steadily decreasing upward force from the moment it leaves the boy's hand until it reaches its highest point; on the way down there is a steadily increasing downward force of gravity as the object gets closer to the earth. c. An almost constant downward force of gravity along with an upward force that steadily decreases until the ball reaches its highest point; on the way down there is only a constant downward force of gravity. d. An almost constant downward force of gravity only e. None of the above. The ball falls back to the ground because of its natural tendency to rest on the surface of the earth.	Q1. A coin is tossed straight upward. What is/are the force(s) that act on the coin after it has been released and as it travels upward? (Ignore air resistance.) a. No forces. b. An upward force only c. A downward force only d. Both an upward and a downward force Q1.3. The coin slows down because the thtotal force decreases as the coin moves upward. a. True b. False Q1.4. The total force stays the same as the coin gets higher. a. True b. False
Q20. A woman exerts a constant horizontal force on a large box. As a result, the box moves across a horizontal floor at a constant speed "V0". The constant force applied by the woman a. has the same magnitude as the weight of the box. b. is greater than the weight of the box. c. has the same magnitude as the total force which resists the weight of the box. d. is greater than the total force which resists the motion of the box. e. is greater than either the weight of the box or the total force which resists its motion.	M18.3 A person is pushing a box across a horizontal floor so that the box moves at a constant speed to the right. There is friction between the box and the floor. M18.3. There is a force pushing the box to the right that is bigger than the friction force acting on the box. a. True b. False

Table 3.9

Comparison of Responses for Aligned MAFA and FCI Items

Group of Interest	Number in Group	Question	Response	Obs	Exp	*df*	X^2
Response *a* to Q21	36	Q1	*d*	18	9	1	12.00***
			a, b, or *c*	18	27		
Response *b* to Q21	131	Q1	*b*	21	32.75	1	5.62*
			a, c, or *d*	110	98.25		
Response *a* to Q1.3	166	Q21	*b*	83	33.2	1	93.38***
			a, c, d, or *e*	83	132.8		
Response *c* to Q21	112	Q1	*d*	65	28	1	65.19***
			a, b, or *c*	47	84		
Response *d* to Q21	49	Q1	*c*	41	8	1	89.97***
			a, b, or *d*	12.25	36.75		
Response *a* to Q18.3	282	Q20	*d* or *e*	167	112.8	1	43.41***
			a, b, or *c*	115	169.2		

*$p < .05$, **$p < .01$, ***$p < .001$

distribution of responses to the paired items differed from what would be expected by chance and these results are also included in Table 3.9. For instance, there were 36 participants who chose answer *a* to Q21. This was the group of interest for the first comparison that is shown in Table 3.9. If there were no relationship between this response and responses to Q1, then one-fourth of participants would be expected to choose each of the four answers to Q1. Because there were 36 participants in the group of interest, nine would be expected to choose answer *d* and the other 27 would be expected to choose one of the other three answers. The observed values show the actual number who chose answer *d* ($n = 18$) and the number that chose one of the other three answers ($n = 18$). The goodness of fit test shows that the number of the participants that chose answer *a* to Q21 that also chose answer *d* to Q1 was significantly higher than that expected by

chance, $X^2 = (1, N = 36) = 12.00$, $p < .001$. A total of six comparisons (including the one described in detail above) are summarized in Table 9. For five of the comparisons a significantly higher number of participants than would be expected by chance chose the aligned answer. For the remaining comparison, a significantly smaller number of participants chose the aligned answer. Overall, the results support the assertion that students responded to the MAFA and the FCI items in similar ways.

Discussion

The purpose of this research was to investigate the efficacy of a new test format for cognitive diagnostic assessments that measure knowledge as a continuous latent construct and misconceptions as a set of discrete skills. This was done by developing a cognitive diagnostic assessment—the Misconceptions About Force Assessment (MAFA)--to measure knowledge and misconceptions about Newton's laws of motion. In the new assessment, misconceptions are modeled with a specific DCM, the *DINA* (*deterministic input noisy-and-gate*) model. A search of the literature revealed three recent DCMs that have been proposed to measure misconceptions. The *Bug-DINO (Bug-diagnostic input noisy or gate)* model (Kuo et al., 2016) is designed to measure only misconceptions. The other two models, the *SISM* (*Simultaneously Identifying Skills and Misconceptions*) model (Kuo et al., 2018) and the *SICM* (*Scaling Individuals and Classifying Misconceptions*) model (Bradshaw & Templin, 2014), are designed to simultaneously measure knowledge and misconceptions. Both measurement models were evaluated through simulation studies and by retrofitting them to existing assessment data. Neither has been evaluated by creating a cognitive diagnostic assessment such as the MAFA and fitting responses from the assessment to the models. In contrast, responses to the FCI, which was

designed to provide a single knowledge score based on CTT, have been evaluated using IRT, DCMs, and other methods to provide measures of knowledge and misconceptions. A few studies have suggested methods to use incorrect answers on the FCI to measure misconceptions (Bao & Reddish, 2001; Fulmer, 2015; Martin-Blas et al., 2010; Saivinainen & Scott, 2002a, 2002b; Savinainen & Viiri, 2008; Yasuda & Taniguchi, 2013), but none of these methods results in a psychometrically sound profile of misconceptions for individuals similar to that provided by the MAFA.

The first research question asked how well the specified measurement models (IRT and DINA) fit responses to the MAFA. Because model fit can be a matter of judgement, I will provide some context for judging the fit of the MAFA by comparing it to the SISM, SICM, and FCI. First, it is useful to consider the structure of the tests. The SISM, SICM and FCI rely on typical multiple-choice questions in which the distractors are aligned with student misconceptions to gather data about knowledge and misconceptions. The MAFA uses a different format. Knowledge is measured using responses to multiple-choice questions (called *knowledge items*) and misconceptions are measured primarily through sets of true/false questions (called *reason items*) that follow each multiple-choice question. Responses to the *knowledge items* were fit to IRT models which were used to estimate overall ability scores for respondents. Responses to the *reason items* were fit to the DINA model and used to estimate the most likely misconception profiles for the respondents. This format allows the identification of misconceptions for students who answer knowledge items correctly. One other model, the SISM, also identified coexisting knowledge and misconceptions. However, the information it provided was different than that provided by the other model and tests. The SISM measured knowledge as a set of discrete skills instead of a continuous latent construct. Rather than providing a profile of

skills and misconceptions, the SISM placed respondents into one of four classes: 1) possesses all skills and no misconceptions, 2) possesses all skills and at least one misconception, 3) missing at least one skill and has no misconceptions, and 4) missing at least one skill and has at least one misconception. Because the misconception profiles provided by the MAFA, SICM, and retrofitted FCI provide more detailed information for tailoring instruction, further comparisons will focus on these models.

The MAFA and SICM model knowledge as a continuous latent construct using IRT. The FCI models knowledge as a continuous latent construct using CTT, but has been retrofitted with IRT models (Planninic et al., 2010; Wang & Bao, 2010). In fact, the SICM was evaluated using a set of 10,039 FCI responses (Bradshaw & Templin, 2014). The measurement models used in all cases were built upon an assumption of unidimensionality. Exploratory factor analysis of the final 12-item version of MAFA knowledge items provided mixed statistical evidence for a unidimensional model with the AIC indicating better fit for a two-dimensional model and BIC for a one-dimensional model. However, the case for the unidimensionality of a test should be based on both statistical analysis and reasonable judgements about the latent constructs measured by the test (Kane, 2006) and, using this guidance, I determined that the one-dimensional model made more sense. Although the MAFA knowledge items ask different questions than the FCI items, they are based upon the same underlying research and use many of the same scenarios. Therefore, it may be useful to compare the dimensionality of MAFA knowledge items to the results of research on the dimensionality of the FCI.

The authors of the FCI claimed that the items represented a single force concept based on 6 underlying dimensions (Hestenes et al., 1992). Researchers who have investigated the fit of multidimensional models to FCI data have presented arguments for multidimensional models

and even varying dimensionalities for data sets from different populations (i.e., high school students versus university students). Huffman and Heller (1995) performed factor analysis on two sets of FCI data—one from high school students and another from university students. For the high school data, 7 of the 30 FCI questions loaded onto two main factors and the remaining 23 items were split between eight factors none of which accounted for a significant amount of variance. For the university data, five items loaded onto a single factor and the remaining 25 items loaded onto eight factors none of which accounted for a significant amount of variance. The authors suggested multiple ways that their findings could be explained. First, students may not conceptualize the questions on the FCI in the same ways that the FCI authors do. For instance, students' explanations for situations might be context dependent. Second, students' knowledge may not be coherently organized around larger underlying principles. In contrast to Huffman and Heller (1995), Scott and Schumayer (2012) performed EFA on an FCI data set and found support for both a one-dimensional model and a five-dimensional model. In both cases, they found that the factors made sense. It should be noted that the authors of the FCI did not necessarily expect students to conceptualize the questions in a manner consistent with a unidimensional structure (Hestenes & Halloun, 1995).

Because the MAFA knowledge items are based on similar scenarios as FCI items, the effect of students' conceptualizations on their responses should be considered. Data on students' interpretations of MAFA items were gathered during the think-alouds that were conducted during Phase 2 of the research. Two of the participants indicated some amount of cognitive dissonance in their thinking. They questioned whether the same rules that apply to objects on Earth applied to the rocket in space. The two knowledge items that asked about the rocket were excluded from the final model and would not have affected any measures of model fit. However,

it is possible that other items were also interpreted in a context specific manner. Knowledge scores from the MAFA, the FCI, and other concept inventories are used in research to measure the effect of instruction on student understanding. Fragmented student knowledge affects the validity of interpreting any concept inventory responses within these unidimensional frameworks, but the practice of using them in this manner is well established.

Of the two IRT models fit to the 12-item version of the MAFA knowledge items, the 2-PL model showed better overall fit. Item level fit indices indicated good fit with SEs for all item parameters less than or equal to 0.26, no significant item-level chi-square values (as calculated using jMetrik), and all chi square values for local dependence less than 10.0. The software used to estimate the final IRT models provided only relative measures of overall model fit. This is consistent with other studies in which FCI data were fitted to an IRT model. For instance, Wang and Bao (2010) fit a 3-PL IRT model to FCI data and showed acceptable item fit but did not report any overall model fit statistics such as RMSEA. Similarly, Planinic et al. (2010) fit a Rasch model to FCI data and provided only item-level fit data. Finally, Bradshaw and Templin (2014), in fitting FCI data to the SICM model, provided only relative measures of overall model fit for ability estimates. Measures of model fit for the knowledge items of the MAFA compare favorably with the results of the three studies listed above.

Responses to the MAFA reason items were fitted to the DINA model. After 11 of the original 38 items were deleted based on their item parameters, overall model fit was moderate with mean RMSEA = 0.076. However, there were still five items that had RMSEA values greater than 0.10. There are not firm guidelines for what constitutes acceptable model fit in the DINA model, but there is information about the fit of other assessments that have been retrofit with the model. Here, I report on the fit of the MAFA reason items in comparison to these other

studies. The estimated slipping (*s*) and guessing parameters (*g*), item-level RMSEA, and item discrimination index values were used as a first measure of model fit. In a series of simulation studies, de la Torre et al. (2010) found that lower values for *s* and *g* resulted in more accurate item parameter estimates and attribute classifications. Rupp et al. (2010) consider items with small guessing and slipping parameters and large item discrimination values to be well performing in the DINA model. Examples of retrofitting DCMs to existing assessments provided no guidance on item selection as they did not delete any of the items and included items with very large slipping and guessing parameters. For instance, George and Robitzsch (2015) fit responses to the Certificate for Proficiency in English to the DINA model included items with guessing parameters greater than 0.75. During the stepwise deletion of items as described in the methods section, it was noted that the order in which items were deleted and the size of the steps used affected which items were left in the final model. If items with large slipping or guessing parameters or small item discriminations were removed after the first step, RMSEA values varied in ways that caused different items to be deleted at the next step. A more extensive investigation of the effect of item deletion on item parameter estimates could help to establish guidelines for the construction of future CDAs.

The deletion of the eleven reason items from the DINA model resulted in the best model fit for the models that were compared, but some fit indices were still greater than desired in the final model. The number of items measuring each misconception in the final model varied widely with 14 items measuring M1, one item measuring M2, six items measuring M3, 11 items measuring M4, two items measuring M5, and four items measuring M6. The large number of items measuring M1 may help to explain the large correlations between M1 and the other five misconceptions. (Values ranged from 0.75 to 0.97.) In a simulation study using the SICM model,

Bradshaw and Templin (2014) suggested that correlations between misconceptions would likely be smaller than 0.7--a typical value found in applications for DCMs that measure skills--and found correlations of just above 0.5 to near zero between three misconceptions using selected FCI data. However, it should be noted that Bradshaw and Templin randomly chose to include the first three of over 30 misconceptions specified by the FCI authors in their model. It is possible that selecting a different set of misconceptions might have given different results. Other possible reasons for the large correlations between M1 and the other misconceptions are that M1 underlies the other misconceptions or that students do not differentiate between M1 and the other misconceptions. Alternatively, large correlations between misconceptions could be due to other factors including misspecification of the Q-matrix, model misspecification in choosing the DINA model instead of a different DCM, and examinee error due to examinees not putting forth maximum effort in responding to the items. A second concern is that there was only one item measuring M2—an object moving in a curved path has an outward force acting on it-- in the final model. One possible reason is that the images included in the four knowledge questions that aligned to this misconception showed the objects as viewed from above. During think-alouds that occurred in Phase 2 of the study, some of the participants indicated confusion about the images although they did end up interpreting as intended. It is possible that some respondents were confused during the pilot and field tests and that this affected the performance of the items and caused them to be deleted.

The MAFA's reason questions are designed to measure student misconceptions. It is assumed that these misconceptions are part of an organized set of ideas that students engage repeatedly when they are asked to make sense of the phenomena presented in the questions. It is possible that that some students' knowledge is so loosely organized that it cannot be measured

with an assessment like the MAFA. Vosniadou and Skopeliti (2014) describe this level of physics understanding as a "framework theory"—a loosely structured set of related concepts that are based on everyday culture and experience and are rooted in a set of ontological beliefs. Framework theories are usually private and held to a lower standard of forecasting power and internal consistency than scientific theories. Students with such a set of beliefs would be unlikely to answer the MAFA reasons items in ways that are consistent with the Q-matrix. This is supported by a reliability study of the FCI in which students took the test twice in a week (Lasry et al., 2011). While students' overall scores were found to change little, on average a third of answers were changed between the test and retest with incorrect answers changed more often than correct answers. There was an 82% chance of choosing a given correct answer both times and the chance of choosing the same incorrect answer twice was only 57%. The authors noted that they did not find their students' inconsistency in choosing incorrect answers surprising given their low average overall FCI score of less than 50%. Some researchers have argued that novice's conceptualizations, although often incorrect, tend to be organized and applied somewhat consistently (diSessa, 1993). The more well organized and consistently applied students' conceptions are, the more effective a concept inventory like the MAFA will be. It is possible that the incorrect answers to MAFA items could be more consistent than those to the FCI because FCI answer choices tend to be longer and include more information. The structure of the MAFA, in which information is spread between multiple items within each set of knowledge and reason items, allows for shorter answer choices which may be clearer to novice thinkers. In summary, a test design like the one implemented in the MAFA can be used to create a CDA for measuring coexisting knowledge and misconceptions. Responses to the MAFA knowledge items showed good fit to the 2-PL model. While responses to most reason items

showed good item level fit, overall model fit was mediocre to poor as measured by the statistics in the CDM package. Overall, the MAFA items showed good fit to the 2-PL IRT model and mediocre to poor fit to the DINA model. It is possible that a different combination of reason items or a different DCM or combination of DCMs might have resulted in better overall model fit.

The second research question concerned how consistent students' responses were between MAFA and FCI items that measured the same misconceptions. The results showed that five of six matched items showed a significantly greater consistency between answers than expected by chance. This consistency provides evidence that the MAFA items and the FCI items measured the same latent constructs. It should be noted that the misconceptions assigned to the FCI questions in this study were different than the misconceptions assigned by the FCI's authors. For instance, the authors of the FCI aligned the item about an item tossed into the air with the misconceptions "impetus dissipation", "gravity intrinsic to mass", and "gravity increases as objects fall" (Hestenes et al., 1992, p.144), none of which were used in the MAFA.

This study contributes to the field of educational measurement by suggesting a different structure for a cognitive diagnostic assessment to measure misconceptions and by demonstrating the development of a new assessment that uses the structure and fitting it with the IRT and DINA models. While multiple studies have suggested that cognitive diagnostic models could be applied to measure misconceptions, to date the emphasis has been on developing new DCMs and retrofitting responses to selected items from existing assessments to them. For instance, responses to selected items from the FCI were fit to the SICM model (Bradshaw & Templin, 2014) and selected responses from a Taiwanese primary school math assessment were fit to the Bug-DINO model (Kuo et al., 2016) and the SISM model (Kuo et al., 2018). This study took a

different approach. Instead of developing a new DCM and fitting it to a traditionally structured concept inventory, it used a different test structure that could be fit with existing measurement models.

This project contributes to Science Education Research by demonstrating a new method to construct a concept inventory that provides a profile of student misconceptions. Understanding what misconceptions students possess can help teachers to modify instruction so that students are less likely to leave school possessing the same misconceptions with which they entered. Tools such as the MAFA may help teachers gauge student misconceptions efficiently. The structure of the MAFA allows the identification of misconceptions about physical situations for all students—both those who answer initial questions about forces and motion correctly and those who answer incorrectly. Traditional science questions can be answered correctly even when students have misconceptions and this may lead teachers to overestimate their students' understanding (Diakidoy & Iordanou, 2003; Mazur, 2009; Schneps & Sadler, 1988). For instance, there is some evidence that students may answer FCI items correctly even when they possess underlying misconceptions (Thornton, et al., 2009). The MAFA allows for the coexistence of knowledge and misconceptions and provides a tool that can be used to measure both. It is possible for researchers to apply the test structure and test development processes demonstrated in this paper to modify or create other concept inventories to provide profiles of misconceptions in other topics.

Limitations

As with any study, there were limitations to this project. First, was the motivation of participants. Because all interactions with participants in Phases 3 and 4 were by email, it is likely that there were some participants who did not put maximum effort into completing the

assessment. Eliminating responses for participants that took less than nine minutes to complete the assessment is unlikely to have eliminated all participants who used less than maximum effort. Second, in determining which items to include in the final IRT and DINA models, I used one of many possible strategies to eliminate poorly performing items. The application of different selection strategies resulted in different reason items being included in the final model. It is possible that other combinations of items might have provided better overall model fit. Third, after deleting items to improve model fit, there was only one reason item that measured M2. Fourth, although the sample size was sufficient to make the IRT models converge, it was not large enough to test random samples of the data for stability of item parameters and for DIF. Fifth is that to achieve a sample size large enough to make the models converge, I combined the pilot test data with the field test data. The pilot test data had higher mean knowledge scores, a greater percentage of male respondents, and a greater percentage of respondents who had completed physics courses—especially advanced physics course—in high school. Finally, the reason items were only tested with the DINA model. It is possible that a different DCM or combination of DCMs may have produced better model fit.

Suggestions for Further Research

One area for further research is to repeat the project with a larger sample from students who are more likely to be motivated to complete the assessment with maximum effort. This might be achieved by asking instructors to administer the assessment to their students as part of their course. A second area is to investigate different rules for keeping or eliminating items from the test. This might help to establish guidelines for construction of CDAs in the future. A third possibility is to compare model fit for the DINA and NIDA models. A fourth possibility is to apply the structure of the MAFA to develop other concept inventories to measure

misconceptions. Finally, it could be possible to research the possibility of presenting items as a computer adaptive test in which an individual's profile is estimated after each item or group of items and the test ends once the estimation is precise.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transaction Automatic Control, 19*(6), 716-723.

Alpir, S., & Duygu, A. (2017). The effects of test length and sample size on item parameters in item response theory. *Education Sciences: Theory and Practice, 17*(1), 321-335.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and Evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*, *39*(10), 952-978. doi:10.1002/tea.10053

Bao, L. & Reddish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *American Journal of Physics, 69*(S1), S45-S53. Retrieved from https://files.eric.ed.gov/fulltext/ED461516.pdf

Bardar, E. M., Prather, E. E., Brecher, K., & Slater, T. F. (2006). Development and validation of the light and spectroscopy concept inventory. *Astronomy Education Review, 5*(2), 103-113. doi:10.3847/AER2009024

Bradshaw, L. & Templin, J. (2014). Combining item-response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403-425. https://doi.org/10.1007/s11336-013-9350-4

Brown, D. E., & Hammer, D. (2013). Conceptual change in physics. In S. Vosniadou (Ed.), *International handbook of research on conceptual change*. New York, NY: Routledge.

Cai, L., Thissen, D., and du Toit, S. H. C. (2017a). IRTPRO for Windows (Version 4.2) [Computer Software]. Lincolnwood, IL: Scientific Software International.

Cai, L., Thissen, D., & du Toit, S. H. C. (2017b). IRTPRO 4.2 Student Guide. Linconwood, IL: Scientific Software International.

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48,* 1074-1079. https://doi.org/10.1119.12290

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152. https://doi.org/10.1207/s15516709cog0502_2

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*(1), 66-71. doi:10.1119/1.12989

Clement, J. (1998). Expert novice similarities and instruction using analogies. *Journal of Science Education, 20*, 1271-1286. doi:10.1080/0950069980201007

Cummings, K., Laws, P. W., Redish, E. F., & Cooney, P. J. (2004). *Understanding physics: Part 2*. United States of America: John Wiley and Sons, Inc.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guildford Press.

de la Torre, J. (2009). A cognitive diagnostic model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163-183. https://doi.org/10.1177/0146621608320523

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*(2), 227-249.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educativa, 20*, 89-97. https://doi.org/10.1016/j.pse.2014.11.001

Diakidoy, I. N., & Iordanou, K. (2003). Preservice teachers' and teachers' conceptions of energy and their ability to predict pupils' level of understanding. *European Journal of Psychology of Education*, *18*(4), 357–368. doi:10.1007/BF03173241

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (361-390). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dick-Perez, M., Luxford, C. J., Windus, T. L., & Holme, T. (2016). A quantum chemistry concept inventory for physical chemistry classes. *Journal of Chemical Education, 93*(4), 605-612. doi:10.1021/acs.jchemed.5b00781

Dillman, D., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail and mixed-mode surveys: The tailored design method, 4th Edition*. Hoboken, NJ: John Wiley and Sons, Inc.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physics Review Special Topics: Physics Education Research, 2*(1), 010105-1-010105-7. https://doi.org/10.1103/PhysRevSTPER.2.010105

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2 & 3), 105-225.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, *20*(10), 1155–1191.

Duit, R. (1993). Research on students' conceptions—developments and trends. *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Misconceptions Trust: Ithaca, NY.

Duit, R., & Treagust, D. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*(6), 671-688. https://doi.org/10.1080/0950069030501 6

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools*, *43*(3), 247–265. https://doi.org/10.1002/pits.20147

Fulmer, G. W. (2015). Validating proposed learning progressions on force and motion using the force concept inventory: Findings from Singapore secondary schools. *International Journal of Science and Mathematics Education, 13*, 1235-1254. https://doi.org/10.1007/s10763-01409553-x

George, A. C., & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology, 11*(3), 189-205. doi:10.20982/tqmp.11.3.p189

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Unlu, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1-24. doi:10.18637/jss.v074.i02

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64-74. https://doi.org/10.1119/1.18809

Halloun, I. A., & Hestenes, D. (1985a). The initial knowledge state of college physics students. *American Journal of Physics, 53,* 1043-1055. https://doi.org/10.1119/1.14030

Halloun, I. A., & Hestenes, D. (1985b). Common sense concepts about motion. *American Journal of Physics, 53,* 1056-1073. https://doi.org/10.1119/1.14031

Harrison, A. G., & Treagust, D. F. (2001). Conceptual change using multiple interpretive perspectives: Two case studies in secondary school chemistry. *Instructional Science, 29*, 45-85. https://doi.org/10.1023/A:1026456101444

Hattie, J. (2015). The applicability of Visible Learning to higher education. Scholarship of Teaching and Learning in Psychology, 1(1), 79-91. http://dx.doi.org/10.1037/stl0000021

Hestenes, D., & Halloun, I. (1995). A response to March 1995 critique by Huffman and Heller. *The Physics Teacher, 33*(11), 502. doi:10.1119/1.2344278

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher, 30*(3), 159-166. https://doi.org/10.1119/1.2343498

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158. https://doi.org/10.1119/1.2343497

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi-org.ezproxy.lib.vt.edu/10.1080/10705519909540118

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher, 33*, 138-143. https://doi.org/10.1119/1.2344171

Jarrett, L., Ferry, B., & Takacs, G. (2012). Development and validation of a concept inventory for introductory-level climate change science. *International Journal of Innovation in Science and Mathematics Education, 20*(2), 25-41.

Jones, M. G., & Brader-Araje, L. (2002). The impact of constructivism on education: Language, discourse, and meaning. *American Communication Journal, 5*(3). Retrieved from https://pdfs.semanticscholar.org/f674/80594ca2ab46e25777653a8cc4f05fbe3135.pdf

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item-response theory. *Applied Psychological Measurement*, *25*(), 258-272.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement, 4th Edition* (17-64). Westport, CT: Praeger Publishers.

Klymkowsky, M. W., Garvin-Doxas, K., and Zeilik, M. (2003). Bioliteracy and teaching efficacy: what biologists can learn from physicists. *Cell Biology Education 2*, 155–161. doi:10.1187/cbe.03-03-0014

Kuo, B., Chen, C., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement, 42*(3), 179-191. doi:10.1177/0146621617722791

Kuo, B., Chen, C., Yang, C., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple choice and constructed-response items. *Educational Psychology, 36*(6), 1115-1133. doi:10.1080/01443410.2016.1166176

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the Force Concept Inventory. *American Journal of Physics, 79*(9), 909-912. doi:10.1119/1.3602073

Lee, Y., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT: An empirical investigation. *Asia Pacific Education Review, 13*, 333-345. doi:10.1007/s12564-011-9196-3

Libarkin, J. C., & Anderson, S. W. (2005). Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, *53*(4), 394–401. https://doi.org/10.5408/1089-9995-53.4.394

LoPresto, M. C., & Murrell, S. R. (2011). An astronomical misconceptions survey. *Journal of College Science Teaching, 40*(5), 14-22. Retrieved from http://www.jstor.org/stable/42993871

Lucariello, J., & Naff, D. (n.d.). How do my students think: Diagnosing student thinking. Retrieved from https://www.apa.org/education/k12/student-thinking

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement: Interdisciplinary Research and Perspectives, 11*, 71-137.

McDonald, R. P., & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

Martin-Blas, T., Seidel, L., & Serrano-Fernandez, A. (2010). Enhancing force concept inventory diagnostics to identify dominant misconceptions in first-year engineering physics. *European Journal of Physics Education, 35*(6), 597-606. https://doi.org/10.1080/03043797.2010.497552

Matthews, M. R. (1998). Introductory comments on philosophy and constructivism in science education. In M. R. Matthews (Ed.), *Constructivism in Science Education* (1-10), Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mazur, E. (2009). Farewell, lecture? *Science, 323*(5910), new series, 50-51. doi: 10.1126/science.1168927

McCloskey, M, Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science, 210,* 1139-1141. doi: 10.1126/science.210.4474.1139

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning Memory and Cognition, 9*(4), 636-649.

McDermott, L.C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, *37*(7), 2-10.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037/0003-066x.50.9.741

Meyer, J. P. (2018). jMetrik (Version 4.1.1) [Computer Software]. Charlottesville, VA: Psychomeasurement Systems, LLC.

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The Physics Teacher, 20*(10), 10-14. doi:10.1119/1.2340924

National Research Council. (1997). *Science teaching reconsidered: A handbook*. Washington, DC: The National Academies Press. https://doi.org/10.17226/5287.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and application of mental appraisal (2nd edition)*. Boston, MA: Pearson Education, Inc.

Pavelich, M., Jenkins, B., Birk, J., Bauer, R., & Krause, S. (2004). Development of a chemistry concept inventory for use in chemistry, materials, and other engineering courses. *Proceedings of the 2004 American Society for Engineering Education Annual Conference,* American Society for Engineering Education. Retrieved from https://peer.asee.org/development-of-a-chemistry-concept-inventory-for.pdf

Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher, 24*(7), 5-12. Retrieved from http://www.jstor.org/stable/1177059

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch Model Based Analysis of the Force Concept Inventory. *Physical Review Special Topics: Physics Education Research, 6,* 010103-1-010103-11. https://doi.org/10.1103/PhysRevSTPER.6.010103

Robitzsch, A., Kiefer, T., George, A. C., & Unlu, A. (2020a). The R package CDM: Cognitive Diagnosis Modeling (Version 7.5-15) [Software]. Available from https://CRAN.R-project.org/package=CDM

Robitzsch, A., Kiefer, T., George, A. C., & Unlu, A. (2020b). CDM: Cognitive Diagnosis Modeling (Version 7.5-15) Manual. Retrieved 1/3/21 from https://cloud.r-project.org/web/packages/CDM/CDM.pdf

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262. doi:10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., & Gould, R. R. (2010). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 National Science Standards. *Astonomy Education Review, 8*(1), 010111-1-010111-26. https://doi-org.ezproxy.lib.vt.edu/10.3847/AER2009024

Savinainen, A., & Scott, P. (2002a). The Force Concept Inventory: A tool for monitoring student learning. *Physics Education, 37*(1), 45-52. doi:10.1088/0031-9120/37/1/306

Savinainen, A., & Scott, P. (2002b). Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education, 37*(1), 53-58. doi:10.1088/0031-9120/37/1/307

Savinainen, A., & Viiri, J. (2008). The Force Concept Inventory as a measure of students' conceptual coherence. *International Journal of Science and Mathematics Education, 6* (4), 719–740. doi:10.1007/s10763-007-9103-x

Schneps, M. H., and Sadler, P. M. (1988). *A private universe* [Motion picture]. Santa Monica, CA: Pyramid Films.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

Scott, T. F., & Schumayer, D. (2012). Exploratory factor analysis of a force concept inventory data set. *Physics Review Special Topics-Physics Education Research, 8*(2), 020105-1-020105-10). Doi:10.1103/PhysRevSTPER.8.020105

Smith, J. I., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*, 115-163.

Smith, J. I., & Tanner, K. (2010). The problem of revealing how students think: Concept inventories and beyond. *CBE-Life Sciences Education, 9*, 1-5. doi:10.1187/cbe.09–12–009

Thornton, R., Kuhl, D., Cummings, K., & Marx, J. (2009). Comparing force and motion conceptual evaluation and the force concept inventory. *Physical Review Special Topics—Physics Education Research, 5*(1), 010105. http://dx.doi.org/10.1103/PhyRevSTPER5.010105

Thornton, R., & Sokoloff, D. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*(4), 338–352.

Unlu, P., & Gok, B. (2007). An investigation of students' misconceptions about centripetal force in uniform circular motion. *Gazi University Journal of Gazi Educational Faculty (GUJGEF), 3*, 141-150. Retrieved from http://login.ezproxy.lib.vt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=30058040&scope=site

Viennot, L. (1979) Spontaneous reasoning in elementary dynamics. *European Journal of Science Education*, *1*(2), 205-221. doi:10.1080/0140528790010209

Viennot, L. (1985). Analysing students' reasoning in science: A pragmatic view of theoretical problems. *European Journal of Science Education, 7*(2), 151-162. https://doi.org/10.1080/0140528850070206

Vosniadou, S. (2014). Examining cognitive development from a conceptual change point of view: The framework theory approach. *European Journal of Developmental Psychology, 11*(6), 645-661. https://doi.org/10.1080/17405629.2014.921153

Vosniadou, S. & Skopeliti, I. (2014). Conceptual change from the framework theory side of the fence. *Science and Education, 23,* 1427-1445. doi:10.1007/s11191-013-9640-3

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics, 78*, 1064-1070. doi: 10.1119/1.3443565

Wenning, C. J. (2008). Dealing more effectively with alternative conceptions in science. *Journal of Physics Teacher education, Online, 5*(1), 11-19.

Williamson, K. (2013). *Development and calibration of a concept inventory to measure introductory college astronomy and physics students' understanding of Newtonian gravity* (Doctoral dissertation). Retrieved from ProQuest.

Williamson, K., Prather, E. E., & Willoughby, S. (2016). Applicability of the Newtonian concept inventory to introductory college physics classes. *American Journal of Physics, 84*(6), 458-466. doi:10.1119/1.4945347

Yasuda, J., & Taniguchi, M. (2013). Validating two questions in the Force Concept Inventory with subquestions. *Physical Science Review Special Topics-Physics Education Research, 9*(1), 010113-1, 010113-7. https://doi.org/10.1103/PhysRevSTPER.9.010113

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 124-145.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.). *Educational Measurement, 4th Edition* (111-154), Westport, CT: Praeger Publishers.

Yeo, S. & Zadnick, M. (2001). Introductory thermal concept evaluation: Assessing students' understanding. *The Physics Teacher*, *39*(11), 496-504.

Chapter 4

The Prevalence and Persistence of Physics Misconceptions

Abstract

The Misconceptions About Force Assessment (MAFA) was used to estimate misconceptions profiles and knowledge scores of 449 undergraduate students who had completed no more than two semesters of university physics. These scores along with the numbers and types of physics courses completed in high school and college were used to investigate the relationship between the possession of misconceptions and two factors: knowledge scores and physics education. One-way ANOVA showed significant difference in knowledge scores between students who possessed each of six misconceptions and those who did not. However, a large proportion of students with high knowledge scores still possessed misconceptions. Logistic regression showed that students who had completed any physics courses were less likely to possess four of the misconceptions than students who had completed no physics courses and that students who had completed an AP or IB course in high school were less likely to possess all six misconceptions. Because only seven respondents who had taken no high school physics courses had completed a college physics course, the performance of this group could not be adequately analyzed. This project provides evidence for the persistence of misconceptions even for students who can answer many physics problems correctly.

Introduction

Understanding science requires more than rote learning. Science is both a "body of knowledge that reflects current understanding of the world" and "a set of practices used to

extend and refine that knowledge" (National Research Council [NRC], 2012, p.26). The most recent national program in the United States to improve K-12 science education and student achievement is the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013). The NGSS are very different from earlier science standards. They are written as a set of performance expectations—three dimensional statements about what students should know and be able to do at different grade levels. Each performance expectation specifies that students will perform a science or engineering practice (SEP) while applying a disciplinary core idea (DCI) and a cross-cutting concept (CCC). Mastery of performance expectations requires students to actively make meaning as they develop a progression of scientific knowledge and skills that are built over the entire course of K-12 education (NRC, 2012). Throughout this process, it is common for students to incorporate incorrect conceptions—misconceptions—into their thinking (NRC, 2012; Posner et al., 1982; Vosniadou & Skopeliti, 2014).

Students enter formal science education with a well-established set of beliefs and personal theories that they use to explain the world some of which may be incorrect (Halloun & Hestenes, 1985a; NRC, 2012). Traditional science education does not eradicate students' misconceptions and even secondary students enter physics instruction with misconceptions that may be difficult to change (Halloun & Hestenes, 1985a). Changing one's existing conceptions requires time. One must be open to new ideas, find a way to incorporate them into one's existing beliefs, and practice applying them to new situations (Montana State University [MSU], 2021). Research on conceptual change shows that the process of learning physics is messy. Students do not simply replace existing misconceptions with correct physics knowledge. New information must be incorporated into an existing framework of ideas and beliefs, some of which contradict scientific thinking (Posner et al., 1982). Instruction often facilitates the development of

misconceptions as students distort the scientific information to fit their existing knowledge (NRC, 2012; Vosniadou & Skopeliti, 2014). The result is a mixture of correct and incorrect scientific knowledge. This allows some students who have misconceptions to appear to understand certain topics because they can solve problems correctly (Posner & Gertzog, 1982).

In physics, misconceptions are resistant to change and often hinder students' progress toward developing a deep understanding of the subject (Halloun & Hestenes, 1985a). High achievement test scores are not always indicators of conceptual understanding—they may be achieved despite the presence of misconceptions (Harrison & Treagust, 2001; Brown & Hammer, 2013). Identifying student misconceptions in physics can be difficult. It is easy for teachers to mistake rote learning for deeper understanding. Even physics students who perform well on tests may retain misconceptions upon completion of a physics course. However, one of the most valuable practices for learning is to provide feedback to students about their misconceptions along with opportunities to correct them (Hattie, 2015). A measure of the prevalence of misconceptions among students of different abilities can help to highlight topics that should most likely be addressed in introductory physics courses.

Currently, there are many concept inventories—multiple choice assessments which measure a set of core knowledge—in physics. Commonly used inventories such as the Force Concept Inventory (FCI) (Hestenes et al., 1992) and the Force and Motion Conceptual Evaluation (Thornton & Sokoloff, 1998) use misconceptions as distractors but do not score them to provide information about the prevalence of specific misconceptions. This study utilizes a new concept inventory—the Misconceptions About Force Assessment (Norris, 2021)—that had been developed to simultaneously measure student knowledge/ability level and identify the presence of six misconceptions about Newton's first and second laws of motion. Both

misconceptions and the physical situations used as the basis for MAFA items were based on prior research into student's misconceptions about forces and motion (Champagne et al., 1980 (in McDermott, 1984); Clement, 1982; Clement, 1998 (as cited in Cummings et al., 2004); McDermott, 1984; Halloun & Hestenes, 1985b; McCloskey et al., 1980; Minstrell, 1982; Unlu & Gok, 2007; Viennot, 1979 (in McDermott, 1984); Wenning, 2008). The use of the MAFA allows me to address two questions:

1. What is the relationship between physics knowledge and the possession of misconceptions about Newton's first and second laws of motion?

2. How is the possession of misconceptions related to the types of physics courses completed?

Method

Participants

Participants were 449 undergraduates who had completed no more than 2 semesters of college level physics at the 200/2000 level. They were recruited through flyers posted on campus, in grocery stores and coffee shops and class announcements in introductory social science, life science, and physical science courses at six public universities in Virginia. Students who were interested in participating contacted me by email and I sent them detailed information about participating along with a link to complete the assessment. Respondents were compensated $5 for completing the assessment to the best of their ability. The time stamp of the survey submission was used as a proxy for effort.

Instrument

The Misconceptions About Force Assessment (MAFA) is a cognitive diagnostic assessment designed to measure knowledge and misconceptions about Newton's first and second laws of motion. The test domain, given in Table 4.1, includes both the knowledge that is measured

Table 4.1

Test Domain for MAFA

Domain for Knowledge Items		
Law	Part	Description
First Law	1.a.	If there are no outside forces acting on an object, it will continue in its state of motion—either at rest or in a straight line at a constant speed.
	1.b.	Objects that are either speeding up or slowing down have a non-zero net force acting on them.
	1.c.	Objects that are moving along a curved path have a non-zero net force acting on them with a component that is perpendicular to the line of motion.
Second Law	2.a.	Objects that are speeding up have a non-zero net force acting on them in the same direction they are moving.
	2.b.	Objects that are slowing down have a non-zero net force acting on them opposite the direction in which they are moving.
	2.c.	The bigger the net force acting on an object, the greater its acceleration.
	2.d.	Objects that are moving in a circle at a constant speed have a non-zero net force acting on them perpendicular to the direction in which they are moving/directed toward the center of the circle.
Domain for Reason Items		
Misconception Number		Description
1		When an object is moving in a given direction, there must be a force acting in that direction.
2		An object moving in a curved path has an outward force acting on it.
3		A constant force causes an object to move with a constant velocity/ an object's velocity is proportional to the magnitude of applied force/changes in speed are caused by changes in the magnitude of applied force.
4		The force of gravity pulls on an object only when it is falling downward.
5		An object that is moving in a curved path will continue to move in a curved path after the removal of the centripetal force.
6		An inanimate or passive object cannot exert a force on a second object because inanimate objects cannot push back.

and expressed as an ability score and the six misconceptions—M1, M2, M3, M4, M5, and M6. The MAFA consists of multiple-choice items (called *knowledge* items) that ask about the forces acting on objects and their resulting motion followed by true/false items (called *reason* items) that ask about reasons for the answers to the multiple-choice items. There are 18 multiple-choice items on the assessment. Each multiple-choice item is followed by between one and three true/false items for a total of 39 true/false items. To improve model fit, not all items were included in the scoring models.

Multiple choice items measure knowledge of Newton's first and second laws of motion which is modeled as a unidimensional construct. Knowledge scores were estimated using the two-parameter logistic (2-PL) item response theory (IRT) model. Because the maximum a posteriori (MAP) estimation was used to estimate item parameters, knowledge scores were calculated for all response patterns including perfect scores. To improve model fit, only 11 items were included in the model. True/false items were aligned to six discrete misconceptions and scores were estimated using the diagnostic inputs noisy-and-gate (DINA) model (Junker & Sijtsma, 2001), a type of diagnostic cognitive model (DCM). Diagnostic cognitive models are designed to measure the possession of a set of discrete skills or attributes. Skills or attributes are typically finer grained than the latent constructs measured by other models. For instance, a construct such as the ability to add fractions could be represented as a set of three smaller skills: 1) adding whole numbers, 2) finding common denominators, and 3) changing improper fractions to mixed numbers. Instead of estimating an overall ability to add fractions, a CDA would classify each respondent as a master (1) or nonmaster (0) of each attribute by placing each respondent into the most likely skills profile. A CDA that measures a attributes will have 2^a possible skills

profiles. For the adding fractions example, the profiles would be 000, 100, 010, 001, 110, 101, 011, and 111. For the MAFA, skills are replaced by misconceptions. In the misconceptions profiles, a 0 represents the absence of the misconception and a 1 represents the possession of the misconception. For instance, a person with the misconceptions profile 101100 most likely possesses three of the six misconceptions—M1, M3, and M4. True-false items are aligned to the misconceptions "required" to choose a given answer. Not all true-false items are aligned to a misconception. Due to this and to improve model fit, only 27 true-false items were included in the DINA model when the skills profiles were estimated.

Data

Data were collected as part of the pilot and field test for the MAFA. The assessment was administered online using the Qualtrics survey platform. The time stamps on the survey were used as a proxy for effort and only responses from students who took at least 9 minutes to complete the assessment were included in the data set. The responses were used to estimate two pieces of information for each respondent: a knowledge score and a profile of misconceptions. Respondents also provided information about the physics courses they had completed in high school and college.

Data Analysis

All data analysis was performed using IBM SPSS Statistics 26. To answer the first research question, one-way ANOVA was used to compare the mean knowledge scores of students who did and did not have each misconception. Welch's procedure was used to account for significant heterogeneity of variance between groups and the Games-Howell post hoc test was used to distinguish which mean differences were significant. To answer the second research question, students were placed into one of six groups based on the physics courses that they had

completed in high school and college. Binomial logistic regression was used to model the probability of having each misconception for students who had completed different types of physics courses compared to students who had completed no physics courses.

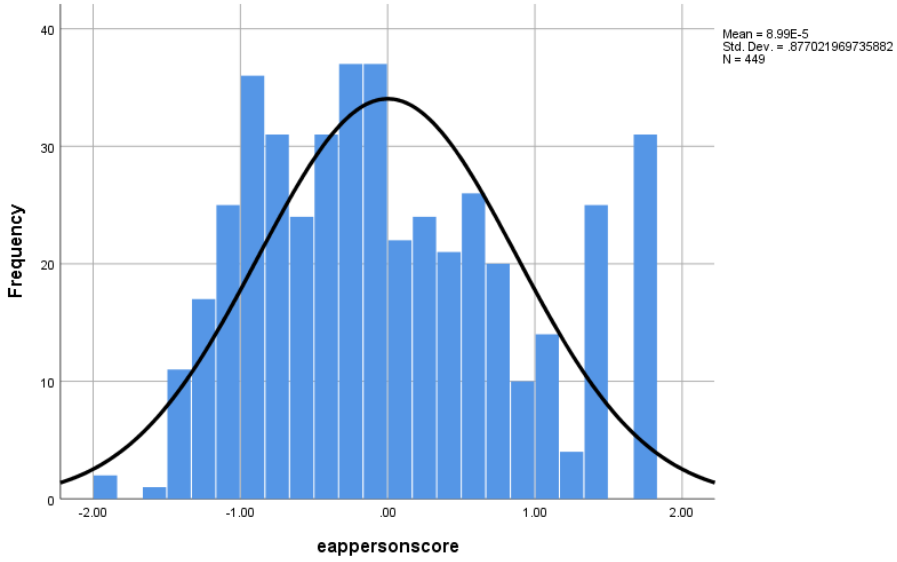
Results

Research Question 1

Knowledge scores (k) ranged between -1.97 and 1.70 with $\bar{k} = 0.00, SD = 0.88$. The IRT model used to estimate the knowledge scores stipulates that the scores fall along a standard normal distribution within the population but does not restrict scores in the sample to this distribution. The distribution of knowledge scores for the data set is shown in Figure 4.1.

Figure 4.1

Distribution of Knowledge Scores in Sample



There were $2^6 = 64$ possible misconceptions profiles for the MAFA, however not every possible profile contained respondents. The number and percent of respondents with each misconception

profile are shown in Table 4.2. In order to conserve space, only profiles that contained respondents

Table 4.2

Respondents in Each Misconception (MC) Profile

Total MCs	MC Profile	Number of Respondents	Percent of Respondents	Cumulative Percent
0	000000	12	2.7	2.7
1	001000	23	5.1	
	000010	43	9.6	
	Total	66	14.7	17.4
2	001010	23	5.1	
	101000	13	2.9	
	100010	2	0.4	
	Total	38	8.5	25.9
3	100011	5	1.1	
	100101	2	0.4	
	101010	25	5.6	
	101100	2	0.4	
	110010	3	0.7	
	Total	37	8.2	34.1
4	101011	10	2.2	
	101101	2	0.4	
	101110	5	1.1	
	110011	1	0.2	
	110101	2	0.4	
	111010	149	33.2	
	Total	169	37.6	71.7
5	101111	18	4.0	
	110111	2	0.4	
	111011	42	9.4	
	111110	21	4.7	
	Total	83	18.5	90.2
6	111111	44	9.8	
	Total	44	9.8	100.0

are included in the table. Most students possessed four or more misconceptions. The most common misconceptions profile of 111010 was assigned to one-third of all students. Three other profiles—000010, 111011, and 111111—account for almost one-tenth of students each. Overall,

77.5% of students had M1, 58.8% had M2, 84.0% had M3, 21.8 had M4, 87.5% had M5, and 28.5% had M6.

One-way ANOVA was performed to compare the mean knowledge scores for students who did and did not have each misconception. Results are shown in Table 4.3. For all six misconceptions, mean knowledge scores were significantly lower for students who possessed the misconception. Difference in knowledge scores between students that did not have the misconception and those that did ranged from 0.77 for M5 to 1.44 for M1. This result is not surprising. Perhaps more interesting are the distributions of knowledge scores for students who did and did not possess each misconception which are shown in Figure 4.2. In these dot plots each

Table 4.3

Difference in Mean Knowledge Score by Misconception

Misconception	Mean knowledge score (SD)			F	df1	df2	p
	No misconception	Misconception	Δk				
1	1.12 (0.58)	-0.32 (0.65)	1.44***	398.52	1	447	0.000
2	0.52 (0.91)	-0.36 (0.64)	0.88***	130.65[a]	1	307.8	0.000
3	0.91 (0.85)	-0.17 (0.77)	1.08***	117.45	1	447	0.000
4	0.21(0.83)	-0.78 (0.50)	0.99***	219.905[a]	1	261.3	0.000
5	0.67 (0.87)	-0.10 (0.84)	0.77***	40.44	1	447	0.000
6	0.25 (0.86)	-0.63 (0.88)	0.88***	167.93[a]	1	366.8	0.000

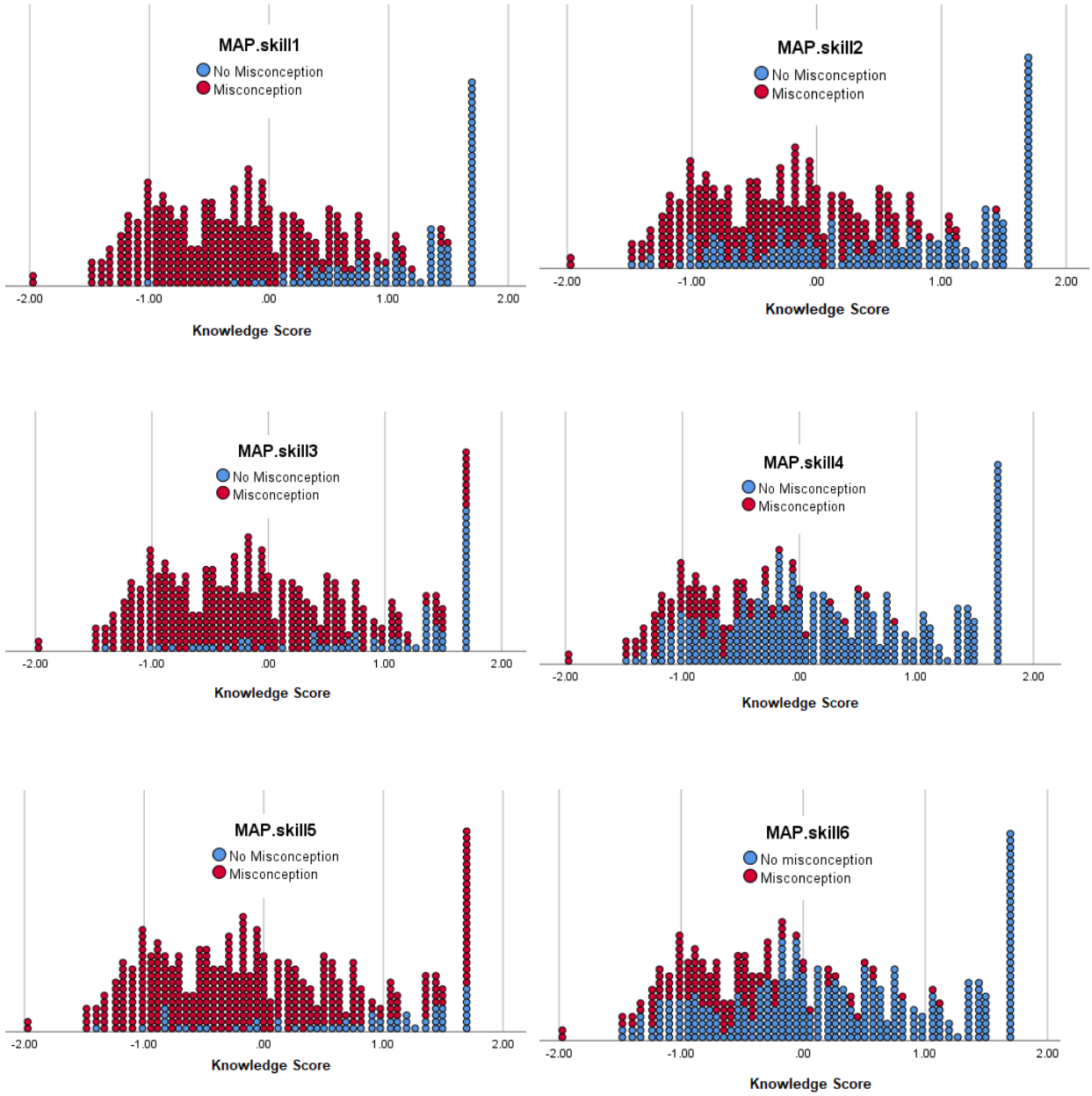
[a] Welch's test statistic reported due to violation of homogeneity of variance assumption.

*** $p < .001$

dot represents one person. From these results it is apparent that students of almost all ability levels possessed the misconceptions.

Figure 4.2

Distribution of Knowledge Scores by Misconception



Research Question 2

Logistic regression was used to investigate differences in the probability of possessing each misconception between groups with different physics education backgrounds. Students were placed into one of six groups based on the physics courses that they had completed in high school and college. For high school courses, conceptual, regular, and honors physics were placed in one group and International Baccalaureate (IB) and Advanced Placement (AP) physics were placed in a second group. Some students had completed multiple physics courses in high school. This was not accounted for in the model. Students were placed into the conceptual/regular/honors group if they had taken only courses in this category and into the AP/IB group if they had taken at least one AP or IB course regardless of other high school physics courses they had completed. The number and percent of respondents in each group are shown in Table 4.4. Only seven respondents who had taken no high school physics had completed a physics course in college. More than one-fourth of students had completed no physics courses. This group was used as the comparison group for the analysis of misconception probability versus physics education.

Table 4.4

Respondents' Physics Education

Physics Courses Completed in College	Types of Courses Completed in High School	Number of Respondents	Percent of Respondents
No physics courses in college	None	123	27.4
	Conceptual/Regular/Honors	155	34.5
	AP or IB	105	23.4
One or two semesters of 200/2000 level physics courses	None	7	1.6
	Conceptual/Regular/Honors	39	8.7
	AP or IB	20	4.5
Total		449	100.0

Table 4.5 shows the results of the logistic regression. For the seven students who had taken only college physics, none had M4 or M6 and all had M5. The regression cannot estimate a meaningful coefficient for these cases, so the coefficients and odds rations for them have been left out of the table. In almost all cases, students who had taken no physics courses were significantly more likely to have each of the misconceptions than students who had taken physics courses. The few exceptions were for students who had taken only conceptual, regular or honors physics courses in high school. These students were just as likely to have M3 and M4 as students who had taken no physics courses and just as likely to have M4 even if they had also completed at least one semester of physics in college. For all six misconceptions, the reduction in odds ratios was smallest for students who had taken only conceptual, regular, or honors physics. For M1, M3, M5, and M6, the odds ratios are similar for students who had taken completed an AP/IB course in high school whether or not they had also completed a college physics course.

Discussion

This study compared the probability of students possessing six misconceptions for students with different knowledge scores and different levels of physics education. Students with higher knowledge scores were less likely to possess all six misconceptions. However, even students with high knowledge scores had misconceptions. While students who completed physics courses-particularly more advanced physics courses—were less likely to have misconceptions than those who had not, students with all levels of education still possessed misconceptions. Large scale international comparisons of student achievement in science and mathematics such as the Programme for International Student Assessment (PISA), given every three years, and the Trends in International Mathematics and Science Study (TIMSS), given every four years, are

Table 4.5

Regression Coefficients for Binomial Regression

MC	Physics Education	B (SE)	Wald	p	OR	95% CI OR
1	None	2.97 (.42)***	50.36	.000	19.50	---
	HS Conc/Reg/Hon	-1.00 (.49)*	4.27	.039	.37	[.14, .95]
	HS AP/IB	-2.80 (.46)***	36.67	.000	.06	[.03, .15]
	College only	-2.68 (.87)**	9.49	.002	.07	[.01, .38]
	HS Conc/Reg/Hon + College	-2.39 (.54)***	19.94	.000	.09	[.03, .26]
	HS AP/IB + College	-3.17 (.61)***	26.66	.000	.04	[.01, .14]
2	None	1.32 (.22)***	35.54	.000	3.73	---
	HS Conc/Reg/Hon	-.72 (.28)*	6.71	.010	.49	[.28, .84]
	HS AP/IB	-1.64 (.30)***	14.24	.000	.19	[.11, .35]
	College only	-2.23 (.87)*	6.66	.010	.11	[.02, .59]
	HS Conc/Reg/Hon + College	-1.47 (.39)***	14.24	.000	.23	[.11, .49]
	HS AP/IB + College	-3.05 (.66)***	21.11	.000	.05	[.01, .17]
3	None	2.97 (.42)***	50.36	.000	19.50	
	HS Conc/Reg/Hon	-.88 (.49)	3.18	.074	.42	[.16, 1.09]
	HS AP/IB	-2.01 (.47)***	18.07	.000	.13	[.05, .34]
	College only	-2.05 (.94)*	4.82	.028	.13	[.02, .81]
	HS Conc/Reg/Hon + College	-2.16 (.54)***	15.78	.000	.12	[.04, .34]
	HS AP/IB + College	-2.12 (.64)**	10.91	.001	.12	[.03, .42]
4	None	-.73 (.19)***	14.38	.000	.48	
	HS Conc/Reg/Hon	-.50 (.27)	3.41	.065	.61	[.35, 1.03]
	HS AP/IB	-1.32 (.36)***	13.24	.000	.27	[.13, .54]
	College only	---	.000	.999	---	[.00, --]
	HS Conc/Reg/Hon + College	-.34 (.41)	.65	.42	.72	[.32, 1.61]
	HS AP/IB + College	-2.21 (1.04)*	4.50	.034	.11	[.01, .85]
5	None	3.39 (.51)***	44.55	.000	29.75	
	HS Conc/Reg/Hon	-1.54 (.56)**	7.56	.006	.21	[.07, .64]
	HS AP/IB	-1.82 (.57)**	10.15	.001	.16	[.05, .50]
	College only	---	.000	.999	---	[.000, ---]
	HS Conc/Reg/Hon + College	-1.69 (.68)*	6.26	.012	.19	[.05, .69]
	HS AP/IB + College	-2.77 (.69)***	16.09	.000	.06	[.02, .24]
6	None	-.32 (.18)	2.91	.088	.73	
	HS Conc/Reg/Hon	-.46 (.25)	3.35	.067	.63	[.39, 1.03]
	HS AP/IB	-1.48 (.33)***	19.73	.000	.23	[.12, .44]
	College only	---	.000	.999	---	[.00, ---]
	HS Conc/Reg/Hon + College	-.89 (.42)*	4.48	.034	.41	[.18, .94]

MC	Physics Education	B (SE)	Wald	p	OR	95% CI OR
	HS AP/IB + College	-1.42 (.65)*	4.76	.029	.24	[.07, .87]

*$p < .05$ ** $p < .01$ *** $p < .001$

designed to compare achievement in science and other disciplines across different countries. The United States has participated in every administration of the PISA which assesses knowledge among students at age 15. Items on the assessment are designed to measure how well students can apply knowledge and skills they have learned in school rather than rote learning—the type of conceptual understanding that is needed for a prepared workforce (Organisation for Economic Cooperation and Development [OECD], 2018). In 2015, seventy-two countries participated in the assessment (OECD, 2018). Results showed the U.S. had only average performance in science—a performance level that has been relatively stable since the Programme's inception (OECD, 2016). Another measure, the Trends in International Mathematics and Science Study (TIMSS) Advanced, measures math and physics achievement among students in their final year of secondary school. In 2015, nine countries participated in the physics portion and the US scored about average—scoring higher than three countries, lower than four, and the same as one other country (Provasnik et al., 2016). Compared to their international peers, secondary students in the United States show average performance in physics.

Average performance in physics by US students may be related to multiple factors. First, the percent of students who complete a high school physics course tends to be low. In 2013 only 39% of US high school students had completed at least one course in physics at graduation (Meltzer et al., 2012). A second reason for the average performance of US students may be that secondary physics courses in the United States tend to be less demanding than those in many other countries. TIMSS Advanced reports the coverage rate—the percent of 18-year-olds who have either taken or are enrolled in a physics course that covers a set list of topics--for each

participating education system. In 2015, the only types of US courses that adequately covered the topics were AP, IB, or second-year physics courses. This put the US coverage rate at 4.8%--second only to Lebanon. In comparison, France, Italy, and Slovenia have coverage rates of 21.5%, 18.2%, and 14.3% respectively (Provasnik et al., 2016). Fewer students in the United States complete rigorous secondary physics courses than in other developed nations despite a ninefold increase in the number of US high school students who take an AP or second course in physics in the last two decades (Meltzer et al., 2012). Changing misconceptions requires engagement with correct conceptions over time. Ideally, students who complete higher level physics courses have more time to form accurate conceptions of the physical world. A third challenge for the United States is a lack of well-prepared secondary physics teachers. Many high school physics courses are taught by teachers who lack an appropriate physics education or experiential background (Meltzer et al., 2012). Teachers who have a tenuous understanding of physics concepts themselves are unlikely to be able to identify and alleviate their student's misconceptions.

The best predictor of student achievement in STEM courses is having a teacher who is certified and has a degree in the field (National Academy of Sciences et al., 2007). Even though fewer than half of high school students take physics, the supply of physics teachers with a degree in the field has not kept up with physics enrollments. According to a report from the Task Force on Teacher Education in Physics (T-TEP), there is a severe, long-term shortage of qualified physics teachers in the US (Meltzer et al., 2012). This shortage poses a great challenge to offering students who enroll in physics a quality physics education. According to the most recent report from the American Institute of Physics, only 39% of recent high school graduates had taken at least one course in physics and only 47% of those courses were taught by a teacher with

either a physics or physics education degree. For comparison, 73% of biology courses and 80% of humanities courses were taught by an educator with a degree in the field (Meltzer et al., 2012). The large number of students who complete a physics course and continue to have misconceptions about the physical world may be impacted by this lack of preparation in the teaching force. The authors of the NGSS acknowledge that one of the greatest challenges of implementing the NGSS is that teachers must undergo conceptual change in their own understanding of what it means to do and teach science before they can lead students to a new understanding (NRC, 2012).

The results of this study agree with earlier research that states that misconceptions in physics are common and resistant to change (Halloun & Hestenes, 1985a; NRC, 2012). Most students who completed the MAFA—even those with high knowledge scores--had multiple misconceptions. Many students who were able to answer questions about the numbers and types of forces acting on objects and the object's motion correctly failed to answer conceptual questions about the reasons for the forces and motion correctly. Students who had completed only conceptual, regular, or honors physics courses in high school had the smallest reduction in the probability of possessing misconceptions compared to students who had completed other courses. Possible reasons for this are that these courses are less demanding, that students are less engaged with the content of the courses, and that teachers of these courses may tend to have a weaker physics background than teachers of more demanding physics courses. Although students who had completed physics courses—particularly AP or IB courses in high school or college courses—were less likely to possess misconceptions, many of these students still had misconceptions.

Finally, the two least common misconceptions, M4 and M6, had the smallest reductions in odds ratios for students who had taken only conceptual, regular, or honors physics. For M4, the odds ratio was 0.48 for students who had taken no physics, 0.29 for students who had taken only conceptual/regular/honors courses in high school, and 0.34 for students who had also completed a college course. For M6, the odds ratio was 0.73 for students who had taken no physics, 0.46 for students who had taken only conceptual/regular/honors courses in high school, and 0.30 for students who had also completed a college course. It could be that less common misconceptions are addressed less frequently in instruction or that students find these misconceptions particularly useful for explaining phenomena and are unlikely to change them.

Limitations of the study include that there was no control for the number of physics courses completed by students, the types of college physics courses completed (i.e. engineering physics versus physics for liberal arts majors), the performance of students in their coursework, or for the quality of teaching students received in the courses they did complete. Each of these factors is likely to affect the probability of possessing misconceptions. In addition, the data were gathered virtually and some respondents may not have put forth maximum effort in completing the assessment. Further research might control for these and other factors to determine the extent to which they affect misconception possession. Future studies might also gather data to compare the effects of completing only a college physics course versus completing a high school course and a college physics course.

References

Brown, D. E., & Hammer, D. (2013). Conceptual change in physics. In S. Vosniadou (Ed.), *International handbook of research on conceptual change*. New York, NY: Routledge.

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48,* 1074-1079. https://doi.org/10.1119.12290

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*(1), 66-71. doi:10.1119/1.12989

Clement, J. (1998). Expert novice similarities and instruction using analogies. *Journal of Science Education, 20*, 1271-1286. doi:10.1080/0950069980201007

Cummings, K., Laws, P. W., Redish, E. F., & Cooney, P. J. (2004). *Understanding physics: Part 2*. United States of America: John Wiley and Sons, Inc.

Halloun, I. A., & Hestenes, D. (1985a). The initial knowledge state of college physics students. American Journal of Physics, 53, 1043-1055. https://doi.org/10.1119/1.14030

Halloun, I. A., & Hestenes, D. (1985b). Common sense concepts about motion. American Journal of Physics, 53, 1056-1073. https://doi.org/10.1119/1.14031

Harrison, A. G., & Treagust, D. F. (2001). Conceptual change using multiple interpretive perspectives: Two case studies in secondary school chemistry. Instructional Science, 29, 45-85. https://doi.org/10.1023/A:1026456101444

Hattie, J. (2015). The applicability of Visible Learning to higher education. Scholarship of Teaching and Learning in Psychology, 1(1), 79-91. http://dx.doi.org/10.1037/stl0000021

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158. https://doi.org/10.1119/1.2343497

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item-response theory. Applied Psychological Measurement, 25(), 258-272.

McCloskey, M, Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science, 210,* 1139-1141. doi: 10.1126/science.210.4474.1139

McDermott, L.C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, *37*(7), 2-10.

Meltzer, D. E., Plisch, M., & Vokos, S. (Eds.). (2012). Transforming the Preparation of Physics Teachers: A call to action. A report by the task force on teacher education in physics (T-TEP). American Physical Society, College Park, MD. Retrieved from https://www.aps.org/about/governance/task-force/upload/ttep-synopsis.pdf

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The Physics Teacher, 20*(10), 10-14. doi:10.1119/1.2340924

Montana State University. (n.d.). *Becoming a 3 dimensional teacher of science: Conceptual change.* Conceptual Change. Retrieved April 4, 2021, from https://www.montana.edu/msse/Framework_Toolkit/conceptual_change.html

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* Washington, DC: The National Academies Press. https://doi.org/10.17226/11463

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). *The Next Generation Science Standards: For state, by states*. Washington, DC: The National Academies Press.

Norris, M. A. (2021). *Rethinking the Force Concept Inventory: Developing a Cognitive Diagnostic Assessment to Measure Misconceptions in Newton's Laws* [Unpublished doctoral dissertation]. Virginia Polytechnic Institute and State University.

Organisation for Economic Cooperation and Development. (2018). PISA 2015 Results in Focus. PISA, OECD Publishing: Paris. Retrieved from http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Organisation for Economic Cooperation and Development. (2016). PISA 2015 Results (Volume I): Excellence and Equity in Education. PISA, OECD Publishing: Paris. https://doi.org/10.1787/9789264266490-en.

Posner, G. J., & Gertzog, W. A. (1982). The clinical interview and the measurement of conceptual change. Science Education, 66(2), 195-209. https://doi.org/10.1002/sce.3730660206

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66(2), 211-227. https://doi.org/10.1002/sce.3730660207

Provasnik, S., Malley, L., Stephens, M., Landeros, K., Perkins, R., & Tang, J. H. (2016). *Highlights from TIMSS and TIMSS Advanced 2015 (NCES 2017-002),* National Center for Education Statistics, Institute for Education Sciences, Department of Education: Washington, DC. Retrieved November 14, 2017, from http://nces.ed.gov/pubs2017/2017002.pdf

Thornton, R., & Sokoloff, D. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*(4), 338–352.

Unlu, P., & Gok, B. (2007). An investigation of students' misconceptions about centripetal force in uniform circular motion. *Gazi University Journal of Gazi Educational Faculty (GUJGEF), 3*, 141-150. Retrieved from http://login.ezproxy.lib.vt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=30058040&scope=site

Viennot, L. (1979) Spontaneous reasoning in elementary dynamics. *European Journal of Science Education*, *1*(2), 205-221. doi:10.1080/0140528790010209

Vosniadou, S. & Skopeliti, I. (2014). Conceptual change from the framework theory side of the fence. Science and Education, 23, 1427-1445. doi:10.1007/s11191-013-9640-3

Wenning, C. J. (2008). Dealing more effectively with alternative conceptions in science. *Journal of Physics Teacher education, Online, 5*(1), 11-19.

Chapter 5

Conclusions

For this dissertation research, I proposed a new format for diagnostic cognitive assessments that measure knowledge and misconceptions. I provided proof of the concept by developing the *Misconceptions About Force Assessment* (*MAFA*)--an assessment for simultaneously measuring knowledge of and diagnosing misconceptions about Newton's first and second laws of motion. I gathered and reported on data related to the validity of the assessment and I used responses to the MAFA to investigate the relationship between students' knowledge and the possession of misconceptions. The research addressed three questions:

1. How well do the specified psychometric models (IRT and DCM) fit responses to the instrument?
2. How do responses on this instrument compare to responses to FCI items which measure the same knowledge and misconceptions?
3. What is the relationship between knowledge (ability) and the presence of specific misconceptions as measured by this instrument?

The methods and research questions were presented in the form of two manuscripts. Chapter 3 consists of the first manuscript which focuses on the test development process and addresses the first two research questions. Chapter 4 consists of the second manuscript which addresses the third research question. In this final chapter I review the main findings of the two manuscripts, describe how they fit within the context of the literature described in Chapters 1 and 2, review

possible future directions for research that were identified, and outline the overall limitations of the study.

Review of Main Findings

The main findings of the study can be divided into four areas. First are findings about the CDA test format that was proposed and demonstrated. Second are findings regarding the application of general test development principles and processes to the development of a diagnostic cognitive assessment. Third are findings related to the types of information provided by the MAFA. The fourth area is findings related to the relationship between knowledge level, physics education, and the presence of misconceptions about Newton's laws. Next, I review the findings and discuss their implications in the same order as they were listed above.

This study presented a test format for a CDA that allows for the simultaneous measure of knowledge and misconceptions. Previous studies approached the application of DCMs to measuring misconceptions by developing new, more complex DCMs that could be used to model responses to typical multiple-choice items (Bradshaw & Templin, 2014; Kuo et al., 2018; Kuo et al., 2016). In addition to their added complexity, each of the new measurement models had limitations. For instance, the *Bug-diagnostic input noisy or gate* (*Bug-DINO*) model (Kuo et al. 2016) measured misconceptions, but not knowledge. The *Simultaneously Identifying Skills and Misconceptions* (*SISM*) model (Kuo et al., 2018) failed to provide complete information about the possession of specific misconceptions. It merely placed respondents into one of two categories based on whether they had no misconceptions or at least one misconception. Finally, the *Scaling Individuals and Classifying Misconceptions* (SICM) model (Bradshaw & Templin, 2014) did not allow for the identification of misconceptions by respondents who chose correct

answers. The limitations of the three proposed models make them less useful as models upon which to base assessments to be used formatively or in a research setting. Simply knowing whether students have at least one misconception will be of little help to designing instruction that targets problematic thinking or to understanding the effects of instruction on the possession of misconceptions. More fine-grained information, such as the presence of specific misconceptions, would be a greater aid to both of these tasks. In addition, it is important to be able to identify misconceptions for students who answer typical test items correctly. Students who perform well on traditional assessments may still possess misconceptions (Brown & Hammer, 2013; Harrison & Treagust, 2001; Schneps & Sadler, 1988). The CDA format proposed here eliminates the limitations of the previous models. It allows for the simultaneous measure of knowledge and diagnosis of specific misconceptions as well as the diagnosis of misconceptions for students who answer questions about topics correctly.

This project approached the problem of diagnosing student's misconceptions differently from previous studies. Instead of creating a new, more complex measurement model, it proposed a new, more complex test structure that used separate sets of test items to measure knowledge and misconceptions. This allowed responses to the items to be fit with existing measurement models--the two-parameter logistic (2-PL) model and the diagnostic inputs noisy-and-gate (DINA) model. The 2-PL model provides an estimate of individual student's knowledge and the DINA model places students into their most probable specific misconception profile. An assessment that diagnoses specific misconceptions is important because all students come to science instruction with preconceptions—facts and mental models--that they use to make sense of the world (Halloun & Hestenes, 1985a; NRC, 2012). Some of their preconceptions are correct and can be used as anchors around which to construct scientifically correct ideas (Lucariello &

Naff, n.d.). However, most students also enter science instruction with misconceptions, ideas that do not agree with scientific theories and facts and their misconceptions can be resistant to change (Halloun & Hestenes, 1985a; NRC, 2012). When misconceptions are included in the construction of new ideas students end up with new misconceptions (NRC, 2012; Posner et al., 1982; Vosniadou & Skopeliti, 2014). The process of restructuring existing mental models to incorporate new knowledge is referred to as conceptual change (DiSessa & Sherin, 1998; Duit & Treagust, 2003). For the conceptual change process to lead to scientifically correct knowledge, students must engage their misconceptions and actively reshape them to produce accurate mental models (NRC, 1997). Knowing which misconceptions students possess can help science educators to design instruction that helps students to meaningfully engage with their misconceptions. Similarly, being able to measure the presence or absence of specific misconceptions can help science education researchers to better understand the extent to which instructional interventions affect their persistence. Both purposes can be aided by assessments such as the MAFA.

Despite taking a different approach to using DCMs to measure misconceptions, this study fills a gap in the literature by providing proof of concept. While Bradshaw and Templin (2014) and Kuo et al. (2018; 2016) suggested that DCMs could be used to diagnose student's science misconceptions, a search of the literature did not reveal any CDAs that did so. The development of a CDA requires the use of more complex psychometric models than those that are frequently used in small-scale assessments as well as knowledge of the topics being tested and this is likely one reason that no examples of CDAs to measure misconceptions were found (de la Torre, 2009). In addition to showing the feasibility of creating a CDA to measure knowledge and misconceptions, the study described the process of test creation. It is a process that can be

followed to create similar assessments. Because the DINA model that was used in the study is the most widely used DCM and one of the least complex (George et al., 2016), it should be more accessible than other models such as the SICM model (Bradshaw & Templin, 2014) to subject matter experts who aim to use the model to create new assessments. Next, I describe the implications of the study for the development of similar assessments about other topics.

The study demonstrated that a psychometrically sound assessment of student knowledge and misconceptions could be developed using the proposed test format and previous research about student misconceptions. The test development process is described in enough detail that it could be used to develop new assessments that follow the same format. Tests that are designed to identify the prevalence of student misconceptions about a scientific topic are often called *concept inventories*. While there are many existing concept inventories (see Table 2.1), I found none that were based on a multidimensional model—the type of model needed to compute a profile of misconceptions. Typically, concept inventories are designed to measure misconceptions indirectly. They are unidimensional tests composed of multiple-choice items in which the distractors are designed to be chosen by students who possess common misconceptions, but the distractors are not scored. Higher overall scores (more correct answers) indicate better conceptual understanding and the presence of fewer misconceptions. Oftentimes, concept inventories are used as a pre- and post-test with the improvement measured as normalized gain (Hake, 1998; LoPresto & Murrell, 2011; Thornton et al., 2009; Williamson et al., 2016; Yeo & Zadnick, 2001). While normalized gain provides a measure of how student knowledge changed, it does not provide diagnostic information that can be used to plan instruction. Using the process described in the project could make it possible to create new CDAs using the research that was used to create existing concept inventories.

Almost all concept inventories that I found were developed by subject-matter experts using classical test theory (CTT). The mathematics of CTT is simpler than that of IRT or DCMs. This allows item parameters to be estimated and interpreted more easily and with smaller sample sizes. Although the 2-PL and DINA models used in this study are more complex than CTT, they are less complex than other models that have been proposed to diagnose misconceptions. In addition, there are open-source software packages (e.g. The R package CDM: Cognitive Diagnosis Modeling (Robitzsch et al., 2020a)) that can be used to estimate both models. Both the lower complexity of the measurement model and the availability of estimation software may make these models more accessible to subject matter experts who want to develop new concept inventories and help to bridge the gap between measurement research and science education research. However, one thing that makes science educators less likely to use the test format presented here is that there is little guidance on determining model fit for the DINA model. The guidelines that I found were ambiguous. Although I described the choices that I made and the reasoning that I used in developing the MAFA, it was clear that different choices about which items to retain might have also produced a moderately well-fitted model. More research is needed to compare different metrics that can be used to retain or delete items from the test before less ambiguous guidelines can be developed. Until such guidelines are available, it is likely that test developers would work in teams of subject matter experts and measurement specialists to develop new CDAs for measuring misconceptions. Next, I discuss the implications of the types of information provided by the MAFA in relation to science education.

The MAFA provides individual student estimates of knowledge about Newton's first and second laws of motion along with a list of which of six misconceptions students are likely to have. Here, I explain how the MAFA can contribute to more effective student learning and why

this is important. There has been a repeated call to improve US STEM education over the past 40 years with the purpose of preparing skilled workers and knowledgeable citizens for an increasingly technical environment (National Academy of Sciences et al., 2007; NCEE, 1983; NRC, 2012; NSF, 2020). Still, US high school students show only average performance in science compared to their international counterparts (OECD, 2016; Provasnik et al., 2016). Researchers have argued that preparation in physics is a key element of science education (Bessin, 2007; Feierman et al., 2006; White, 2008) yet fewer than half of high school graduates complete at least one course in physics (Meltzer et al., 2012). In addition, high school physics courses in the US tend to cover fewer demanding topics than courses in other industrialized countries (Provasniak et al., 2016) and fewer than half of high school physics courses are taught by a teacher who has a physics or physics education degree (Meltzer et al., 2012). The MAFA (and potentially future CDAs modeled after the MAFA) could eventually serve as a tool to strengthen student preparation in physics. It can provide information about student knowledge to teachers who may lack the knowledge or time to recognize student misconceptions without the tool. It could also be used by individual students to self-assess their mastery about Newton's laws. Because the assessment takes little time--most students who took the online test completed it in 10-20 minutes--and because it can be accessed online, teachers and students could use the resource without taking much time from instruction. Of course, one limitation is that that the MAFA has not been field tested with high school students. The most commonly used concept inventory in physics, the Force Concept Inventory, is used with both high school and college students. It is likely that the MAFA would also be valid for use with high school students, but for now this remains an important area for future research.

While the MAFA and, perhaps, future instruments that follow the same design have the potential to improve science instruction, questions remain about how feedback would best be provided. For instance, what should a score report look like and how would the report differ for different users? Novice teachers, experienced teachers, and students would all likely need different types of feedback about their scores and next instructional steps to take based on them. Given that so few high school physics instructors have a physics degree, it is likely that some novice teachers have the same misconceptions as their students. In this case, the MAFA would serve to identify misconceptions in teachers and students. Resources could be provided for novice instructors and students to help them use MAFA scores most effectively. While expert physics instructors may have the skills to identify misconceptions in their students and even know how to design instruction to counteract the misconceptions, they may not have the time to track individual student's misconceptions. The MAFA could provide this information. Research is needed to investigate how novice and experienced physics instructors and physics students would use feedback from the MAFA to improve instruction and learning.

The fourth area to review and discuss is findings about the relationship between knowledge level, physics education, and misconceptions about Newton's laws as measured by the MAFA. For all six misconceptions, students who did not have the misconception had significantly higher knowledge scores compared to students who did have the misconception. However, many students with high knowledge scores still had misconceptions. This was especially pronounced for misconceptions 3 and 5 for which there were students with the highest knowledge score who had the misconception. These results support earlier findings that show the presence of misconceptions even for high-performing students (Halloun & Hestenes, 1985a; Mazur, 2009; NRC, 2012). With a few exceptions (see Table 4.5), students who had completed

any physics course in high school or college were significantly less likely to have misconceptions compared to students who had completed no physics courses. This is good news as it implies that, in general, physics education is effective in helping some students to overcome misconceptions. Students who had completed only conceptual, regular, or honors physics in high school were more likely to have misconceptions than students who had completed a high school AP or IB course or a university course in physics. Possible reasons for this are that the conceptual/regular/honors courses are less rigorous, that the students in these courses tend to be less engaged, and that the teachers of these course tend to have a weaker physics education. In any case, this implies that high school physics teachers and students might best benefit from using the MAFA as an instructional tool. As described above, this remains an important area for research.

The MAFA is an example of a CDA that can be used to measure knowledge and the presence of misconceptions about a science topic. The test format and test creation process described here can serve as a starting point for the creation of new CDAs about other topics. The background research about student misconceptions upon which existing concept inventories are based can be used in the CDA creation process.

Directions for Future Research

Areas for future research fall into multiple categories. First is research about the MAFA itself. The MAFA was only tested with college students. It would be valuable to collect and analyze responses from high school students as well and compare the model fit between the data sets—especially because results indicate that high school instructors and students would benefit more by using the MAFA. How MAFA results would be used and what this means for how they

should be reported is another area for research. It is possible that different users (e.g. experienced teachers, novice teachers, students, and researchers) would benefit from different score reports. In creating the MAFA, data were fitted to only one DCM, the DINA model. Future researchers might investigate the possibility that a different DCM would provide better model fit. Finally, many, but not all, combinations of items were compared to decide which items would be included in the final version of the MAFA. Future research could investigate different combinations for items. It is possible that a different combination of items would also result in a viable test. A second category of research is about the test format and creation process. Future research involving the creation of other CDAs that use the test format can help to further investigate its usability and to refine the test creation process. A third direction for future research involves determining more definite model fit guidelines for the DINA model as used in this context. Reducing ambiguity in model fit guidelines would increase the likelihood that the DINA model will be used to create future assessments. A fourth direction for future research is to use the MAFA to measure the prevalence of misconceptions in different groups and the effects of instructional interventions misconceptions. It could be used as a single tool or it could be used with and compared to other sources of evidence to further build the validity argument.

Limitations

There were several limitations to this project. First, the MAFA was only tested with college students. Second, it is possible that, even after eliminating responses based on the time stamp, some responses in the data set came from poorly motivated students. Third, was the sample size. Although it was large enough to make the IRT models converge, a larger sample would have allowed the comparison of parameter estimates from random samples to check for

for stability and DIF. Fourth, although the item level fit of the response data to the DINA model was good for most items, the overall model fit was mediocre to poor depending on the fit criteria used. While the MAFA could be used in its current form in low stakes environments, further research and/or modifications should be completed before using it in a high-stakes situation. Finally, when comparing student's knowledge scores and physics education to the possession of misconceptions, there were factors such as performance in physics courses and number of physics courses completed that were not controlled for.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transaction Automatic Control, 19*(6), 716-723.

Alpir, S., & Duygu, A. (2017). The effects of test length and sample size on item parameters in item response theory. *Education Sciences: Theory and Practice, 17*(1), 321-335.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and Evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*, *39*(10), 952-978. doi:10.1002/tea.10053

Bao, L. & Reddish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *American Journal of Physics, 69*(S1), S45-S53. Retrieved from https://files.eric.ed.gov/fulltext/ED461516.pdf

Bardar, E. M., Prather, E. E., Brecher, K., & Slater, T. F. (2006). Development and validation of the light and spectroscopy concept inventory. *Astronomy Education Review, 5*(2), 103-113. doi:10.3847/AER2009024

Bessin, B. (2007). Why physics first? *The Physics Teacher, 45*, 134. doi:10.1119/1.2709666

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (392-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bradshaw, L. & Templin, J. (2014). Combining item-response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403-425. https://doi.org/10.1007/s11336-013-9350-4

Brown, D. E., & Hammer, D. (2013). Conceptual change in physics. In S. Vosniadou (Ed.), *International handbook of research on conceptual change*. New York, NY: Routledge.

Cai, L., Thissen, D., & du Toit, S. H. C. (2017). IRTPRO 4.2 Student Guide. Linconwood, IL: Scientific Software International.

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48,* 1074-1079. https://doi.org/10.1119.12290

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152. https://doi.org/10.1207/s15516709cog0502_2

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*(1), 66-71. doi:10.1119/1.12989

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guildford Press.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362.

de la Torre, J. (2009). A cognitive diagnostic model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163-183. https://doi.org/10.1177/0146621608320523

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353. doi: 10.1007/S11336-008-9063-2

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624. https://doi.org/10.1007/s11336-008-9063-2

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educativa, 20*, 89-97. https://doi.org/10.1016/j.pse.2014.11.001

Dempster A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, *39*(1), 1-38. Retrieved from https://www.jstor.org/stable/2984875

Diakidoy, I. N., & Iordanou, K. (2003). Preservice teachers' and teachers' conceptions of energy and their ability to predict pupils' level of understanding. *European Journal of Psychology of Education*, *18*(4), 357–368. doi:10.1007/BF03173241

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement, 39*(1), 62-79. doi:10.1177/0146621614561315

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (979-1030), Amsterdam, The Netherlands: Elsevier.

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (361-390). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dick-Perez, M., Luxford, C. J., Windus, T. L., & Holme, T. (2016). A quantum chemistry concept inventory for physical chemistry classes. *Journal of Chemical Education, 93*(4), 605-612. doi:10.1021/acs.jchemed.5b00781

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physics Review Special Topics: Physics Education Research, 2*(1), 010105-1-010105-7. https://doi.org/10.1103/PhysRevSTPER.2.010105

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2 & 3), 105-225.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, *20*(10), 1155–1191.

Duit, R. (1993). Research on students' conceptions—developments and trends. *The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics,* Misconceptions Trust: Ithaca, NY.

Duit, R., & Treagust, D. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*(6), 671-688. https://doi.org/10.1080/0905006903050l6

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools*, *43*(3), 247–265. https://doi.org/10.1002/pits.20147

Ewald, G., Hickman, J. B., Hickman, P., & Myers, F. (2005). Physics first: The right-side-up science sequence. *The Physics Teacher,* 43, 319-320. doi:10.1119/1.1903844

Feierman, B., Livanis, O., Riendau, D., Hickman, P., & Blanton, P. (2006). *Physics First: An Informational Guide for Teachers, School Administrators, Parents, Scientists, and the Public*. Retrieved from https://www.aapt.org/Resources/physicsfirst.cfm

Feynman, R. P., Leighton, R. B., & Sands, M. L. (1964). The Feynman lectures on physics: Mainly mechanics, radiation, and heat. Reading, MA: Addison-Wesley.

Foote, C. J., Vermette, P. J., & Battaglia, C. F. (2001). *Constructivist strategies: Meeting Standards and Engaging Adolescent Minds.* New York, NY: Taylor and Francis.

Fulmer, G. W. (2015). Validating proposed learning progressions on force and motion using the force concept inventory: Findings from Singapore secondary schools. *International Journal of Science and Mathematics Education, 13*, 1235-1254. https://doi.org/10.1007/s10763-01409553-x

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Unlu, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1-24. doi:10.18637/jss.v074.i02

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (242-274), New York, NY: Cambridge University Press.

Government Publishing Office. (1958). *Public law 85-864*. Retrieved from https://www.gpo.gov/fdsys/pkg/STATUTE-72/pdf/STATUTE-72-Pg1580.pdf

Gunstone, R. F., & White, R. T. (1981). Understanding of gravity. *Science Education*, *65*(3), 291-299.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64-74. https://doi.org/10.1119/1.18809

Halloun, I. A., & Hestenes, D. (1985a). The initial knowledge state of college physics students. *American Journal of Physics, 53,* 1043-1055. https://doi.org/10.1119/1.14030

Halloun, I. A., & Hestenes, D. (1985b). Common sense concepts about motion. *American Journal of Physics, 53,* 1056-1073. https://doi.org/10.1119/1.14031

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14*(2), 75-96. http://www.jstor.org/stable/1434009

Harrison, A. G., & Treagust, D. F. (2001). Conceptual change using multiple interpretive perspectives: Two case studies in secondary school chemistry. *Instructional Science, 29*, 45-85. https://doi.org/10.1023/A:1026456101444

Hattie, J. (2015). The applicability of Visible Learning to higher education. Scholarship of Teaching and Learning in Psychology, 1(1), 79-91. http://dx.doi.org/10.1037/stl0000021

Hechinger Report. (2011). Timeline: Important dates in US science education history. *The worrying state of science education*. Retrieved from http://hechingerreport.org/timeline-important-dates-in-u-s-science-education-history/

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210.

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher, 30*(3), 159-166. https://doi.org/10.1119/1.2343498

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158. https://doi.org/10.1119/1.2343497

Hewson, P. W. & Thorley, N. R. (1989). The conditions of conceptual change in the classroom. *International Journal of Science Education, 11*(5), 541-553. doi:10.1080/0950069890110506

Jarrett, L., Ferry, B., & Takacs, G. (2012). Development and validation of a concept inventory for introductory-level climate change science. *International Journal of Innovation in Science and Mathematics Education, 20*(2), 25-41.

Jones, M. G., & Brader-Araje, L. (2002). The impact of constructivism on education: Language, discourse, and meaning. *American Communication Journal, 5*(3). Retrieved from https://pdfs.semanticscholar.org/f674/80594ca2ab46e25777653a8cc4f05fbe3135.pdf

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item-response theory. *Applied Psychological Measurement*, *25*(), 258-272.

Klymkowsky, M. W., Garvin-Doxas, K., and Zeilik, M. (2003). Bioliteracy and teaching efficacy: what biologists can learn from physicists. *Cell Biology Education 2*, 155–161. doi:10.1187/cbe.03-03-0014

Kuo, B., Chen, C., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement, 42*(3), 179-191. doi:10.1177/0146621617722791

Kuo, B., Chen, C., Yang, C., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple choice and constructed-response items. *Educational Psychology, 36*(6), 1115-1133. doi:10.1080/01443410.2016.1166176

Lee, Y., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT: An empirical investigation. *Asia Pacific Education Review, 13*, 333-345. doi:10.1007/s12564-011-9196-3

Libarkin, J. C., & Anderson, S. W. (2005). Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, *53*(4), 394–401. https://doi.org/10.5408/1089-9995-53.4.394

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548-564.

LoPresto, M. C., & Murrell, S. R. (2011). An astronomical misconceptions survey. *Journal of College Science Teaching, 40*(5), 14-22. Retrieved from http://www.jstor.org/stable/42993871

Lucariello, J., & Naff, D. (n.d.). How do my students think: Diagnosing student thinking. Retrieved from https://www.apa.org/education/k12/student-thinking

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187-212. https://doi-org.ezproxy.lib.vt.edu/10.1007/BF02294535

Martin-Blas, T., Seidel, L., & Serrano-Fernandez, A. (2010). Enhancing force concept inventory diagnostics to identify dominant misconceptions in first-year engineering physics.

European Journal of Physics Education, 35(6), 597-606. https://doi.org/10.1080/03043797.2010.497552

Matthews, M. R. (1998). Introductory comments on philosophy and constructivism in science education. In M. R. Matthews (Ed.), *Constructivism in Science Education* (1-10), Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mazur, E. (2009). Farewell, lecture? *Science, 323*(5910), new series, 50-51. doi: 10.1126/science.1168927

McCloskey, M, Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science, 210,* 1139-1141. doi: 10.1126/science.210.4474.1139

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning Memory and Cognition, 9*(4), 636-649.

McDermott, L.C. (1984). Research on conceptual understanding in mechanics. *Physics Today*, *37*(7), 2-10.

Meltzer, D. E., Plisch, M., & Vokos, S. (Eds.). (2012). Transforming the Preparation of Physics Teachers: A call to action. A report by the task force on teacher education in physics (T-TEP). American Physical Society, College Park, MD. Retrieved from https://www.aps.org/about/governance/task-force/upload/ttep-synopsis.pdf

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The Physics Teacher, 20*(10), 10-14. doi:10.1119/1.2340924

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* Washington, DC: The National Academies Press. https://doi.org/10.17226/11463

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for education reform*. Retrieved from https://ww2.ed.gov/pubs/NatAtRisk/intro.html

National Research Council. (1997). *Science teaching reconsidered: A handbook.* Washington, DC: The National Academies Press. https://doi.org/10.17226/5287**.**

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press. https://doi.org/10.17226/13165

National Science Foundation. (2020). *STEM education for the future: A visioning report*. Retrieved from https:// https://www.nsf.gov/ehr/Materials/STEM%20Education%20for%20the%20Future%20-%202020%20Visioning%20Report.pdf

Novick, S., & Nussbaum, J. (1978). Junior high school pupils' understanding of the particulate nature of matter: An interview study. *Science Education, 62*(3), 273-281. https://doi.org/10.1002/sce.3730620303

Organisation for Economic Cooperation and Development. (2018). PISA 2015 Results in Focus. PISA, OECD Publishing: Paris. Retrieved from http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Organisation for Economic Cooperation and Development. (2016). PISA 2015 Results (Volume I): Excellence and Equity in Education. PISA, OECD Publishing: Paris. https://doi.org/10.1787/9789264266490-en.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and application of mental appraisal (2nd edition)*. Boston, MA: Pearson Education, Inc.

Pavelich, M., Jenkins, B., Birk, J., Bauer, R., & Krause, S. (2004). Development of a chemistry concept inventory for use in chemistry, materials, and other engineering courses. *Proceedings of the 2004 American Society for Engineering Education Annual Conference,* American Society for Engineering Education. Retrieved from https://peer.asee.org/development-of-a-chemistry-concept-inventory-for.pdf

Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher, 24*(7), 5-12. Retrieved from http://www.jstor.org/stable/1177059

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch Model Based Analysis of the Force Concept Inventory. *Physical Review Special Topics: Physics Education Research, 6,* 010103-1-010103-11. https://doi.org/10.1103/PhysRevSTPER.6.010103

Posner, G. J., & Gertzog, W. A. (1982). The clinical interview and the measurement of conceptual change. *Science Education, 66*(2), 195-209. https://doi.org/10.1002/sce.3730660206

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227. https://doi.org/10.1002/sce.3730660207

Powell, A. (2007). How Sputnik changed U.S. education. *The Harvard Gazette*. Retrieved from https://news.harvard.edu/gazette/story/2007/10/how-sputnik-changed-u-s-education/

Provasnik, S., Malley, L., Stephens, M., Landeros, K., Perkins, R., & Tang, J. H. (2016). *Highlights from TIMSS and TIMSS Advanced 2015 (NCES 2017-002),* National Center for Education Statistics, Institute for Education Sciences, Department of Education: Washington, DC. Retrieved 11/14/17 from http://nces.ed.gov/pubs2017/2017002.pdf

Robitzsch, A., Kiefer, T., George, A. C., & Unlu, A. (2020). The R package CDM: Cognitive Diagnosis Modeling (Version 7.5-15) [Software]. Available from https://CRAN.R-project.org/package=CDM

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262. doi:10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

Rutherford, F. J., & Ahlgren, A. (1990). *Science for all Americans*. New York, NY: Oxford University Press.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in*

Science Teaching, 35(3), 265-296. https://doi.org/10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., & Gould, R. R. (2010). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 National Science Standards. *Astonomy Education Review, 8*(1), 010111-1-010111-26. https://doi-org.ezproxy.lib.vt.edu/10.3847/AER2009024

Savinainen, A., & Scott, P. (2002a). The Force Concept Inventory: A tool for monitoring student learning. *Physics Education, 37*(1), 45-52. doi:10.1088/0031-9120/37/1/306

Savinainen, A., & Scott, P. (2002b). Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education, 37*(1), 53-58. doi:10.1088/0031-9120/37/1/307

Savinainen, A., & Viiri, J. (2008). The Force Concept Inventory as a measure of students' conceptual coherence. *International Journal of Science and Mathematics Education, 6* (4), 719–740. doi:10.1007/s10763-007-9103-x

Schneps, M. H., and Sadler, P. M. (1988). *A private universe* [Motion picture]. Santa Monica, CA: Pyramid Films.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement, 67*(2), 239–257. doi:10.1177/001316440629202

Smith, J. I., & Tanner, K. (2010). The problem of revealing how students think: Concept inventories and beyond. *CBE-Life Sciences Education, 9*, 1-5. doi:10.1187/cbe.09–12–009

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (1-49). Oxford, England: Wiley.

Talbot III, R. M. (2013). Taking an item-level approach to measuring change with the force and motion conceptual evaluation: An application of item response theory. *School Science and Mathematics, 113*(7), 356-365. doi:10.1111/ssm.12033

Tatsuoka, K. K. (2009). Cognitive assessment: an introduction to the rule space method. New York, NY: Routledge.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnostic models. *Psychological Methods, 11*(3), 287–305. http://dx.doi.org/10.1037/1082-989X.11.3.287

Thornton, R., Kuhl, D., Cummings, K., & Marx, J. (2009). Comparing force and motion conceptual evaluation and the force concept inventory. *Physical Review Special Topics—Physics Education Research, 5*(1), 010105. http://dx.doi.org/10.1103/PhyRevSTPER5.010105

Thornton, R., & Sokoloff, D. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*(4), 338–352.

Trowbridge, D. E., & McDermott, L. C. (1980). An investigation of student understanding of the concept of velocity in one dimension. *American Journal of Physics*, *48*(12), 1020-1028. https://doi.org/10.1119/1.12298

Trowbridge, D. E., & McDermott, L. C. (1981). Investigation of student understanding of the concept of acceleration in one dimension. *American Journal of Physics*, *49*(3), 242-253. doi:10.1119/1.12525

Viennot, L. (1979) Spontaneous reasoning in elementary dynamics. *European Journal of Science Education*, *1*(2), 205-221. doi:10.1080/0140528790010209

Viennot, L. (1985). Analysing students' reasoning in science: A pragmatic view of theoretical problems. *European Journal of Science Education, 7*(2), 151-162. https://doi.org/10.1080/0140528850070206

von Davier, M. (2005). *A general diagnostic model applied to language testing data (Research Report No. RR-05-16)*. Princeton, NJ: Educational Testing Service.

von Davier, M. (2007). *Hierarchical general diagnostic model (Research Report No. RR-07-19)*. Princeton, NJ: Educational Testing Service.

Vosniadou, S. (2014). Examining cognitive development from a conceptual change point of view: The framework theory approach. *European Journal of Developmental Psychology, 11*(6), 645-661. https://doi.org/10.1080/17405629.2014.921153

Vosniadou, S. & Skopeliti, I. (2014). Conceptual change from the framework theory side of the fence. *Science and Education, 23,* 1427-1445. doi:10.1007/s11191-013-9640-3

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics, 78*, 1064-1070. doi: 10.1119/1.3443565

White, J. W. (2008). Physics first and physics for all (Well sort of). *The Physics Teacher, 46*, 255-256. doi:10.1119/1.2895691

Williamson, K. (2013). *Development and calibration of a concept inventory to measure introductory college astronomy and physics students' understanding of Newtonian gravity* (Doctoral dissertation). Retrieved from ProQuest.

Williamson, K., Prather, E. E., & Willoughby, S. (2016). Applicability of the Newtonian concept inventory to introductory college physics classes. *American Journal of Physics, 84*(6), 458-466. doi:10.1119/1.4945347

Yasuda, J., & Taniguchi, M. (2013). Validating two questions in the Force Concept Inventory with subquestions. *Physical Science Review Special Topics-Physics Education Research, 9*(1), 010113-1, 010113-7. https://doi.org/10.1103/PhysRevSTPER.9.010113

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.). *Educational Measurement, 4th Edition* (111-154), Westport, CT: Praeger Publishers.

Yeo, S. & Zadnick, M. (2001). Introductory thermal concept evaluation: Assessing students' understanding. *The Physics Teacher*, *39*(11), 496-504.

Appendix A

Final Version of Misconceptions About Force Assessment

Default Question Block

Thank you for your interest in this study.

Because many of the questions on this assessment use pictures, we suggest that you use a tablet, laptop, or desktop computer (not a phone) to access the questions.

Please read the consent form below. It contains information to help you decide whether to participate and how to contact the researchers if needed.

RESEARCH SUBJECT CONSENT FORM

Title:	Developing an Assessment to Diagnose Physics Misconceptions—Phase 1
Protocol No.:	VT IRB # 18-917
Sponsor:	Virginia Tech School of Education
Investigator:	Gary Skaggs (Principal Investigator) and Mary Norris Virginia Tech School of Education

1750 Kraft Drive, Room 2104
Blacksburg, VA, 24061
USA

Daytime Phone Number: 540-239-0593

You are being invited to take part in a research study. Participation is voluntary. You can choose not to take part, or agree to take part and later change your mind. There will be no penalty or loss of benefits to which you are otherwise entitled.

The purpose of this research is to ask you questions and determine your feedback. Your participation in this research will last until you have completed the questionnaire. The only risk is effort involved in the questionnaire. There are no benefits to you from your taking part in this research. Others may benefit from the information gained during this research. Your alternative is to not take part in the research. We may publish the results of this research. As we are not collecting any identifiable information, your information will be confidential.

If you have questions, concerns, or complaints, or think this research has hurt you, talk to the research team at the phone number listed above. This research is being overseen by the Virginia Tech Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may talk to them at (540) 231-3732, irb@vt.edu if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

For taking part in this research, you may be paid up to a total of $5.

By continuing in the survey, you are consenting to continue.

This questions on this page ask for demographic data which will be used to test how the items on this assessment perform for different groups of students.

What is your age?

- ◯ 18 years or older (You must be at least 18 years old to participate.)
- ◯ Younger than 18 years

What year are you in school?

- ◯ Freshman
- ◯ Sophomore
- ◯ Junior
- ◯ Senior
- ◯ Prefer not to answer
- ◯ Graduate Student (You must be an undergraduate student to participate.)

What is your sex?

- ◯ Male
- ◯ Female
- ◯ Intersex

◯ Prefer not to answer

Which of the following high school physics classes did you complete?

☐ I completed no physics courses in high school.

☐ Physics First

☐ Conceptual Physics

☐ Regular Physics

☐ Honors Physics

☐ AP Physics 1 (Mechanics, Energy, Waves, and Circuits)

☐ AP Physics 2 (Fluids, Thermodynamics, Electricity and Magnetism, and Atomic & Nuclear Physics)

☐ AP Physics C-Mechanics

☐ AP Physics C--Electricity & Magnetism

☐ IB Physics SL

☐ IB Physics HL

☐ Other

If you answered "Other", please state the title of the course(s) below:

How many ***semesters*** of high school physics did you complete?

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5 or more

How many semesters of your high school physics courses were dual-enrolled? (Dual enrolled courses are those for which you receive credit through a community college without having to take an AP or IB exam.)

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5 or more

How many semesters (**not semester *hours***) of 200 or 2000 level or higher college physics courses have you ***completed***? Please do not include courses in which you are currently enrolled. (Note that you must have completed no more than two semesters of these courses to participate.)

- ○ 0
- ○ 1
- ○ 2
- ○ 3 or more

Please enter the names or course numbers of the 2000 level or higher college physics courses you have ***completed***.

If you are enrolled in a physics course ***this semester***, please list the name or course number here. Otherwise, type "None".

Please enter the name of the university you are currently attending.

Instructions: This assessment consists of 19 multiple-choice questions each of which is followed by 2-4 true/false questions. Each page of the assessment will present either a single multiple-choice question, or a set of 2-4 true/false questions. Each multiple-choice question and the true/false questions that follow it ask about the same situation. Sometimes, the true/false questions that are presented will differ depending on your answer to the multiple-choice question.

Please choose the ***single best answer*** for each question.

Use the forward and back arrows at the bottom of each page to navigate between pages. You can return to previous questions and change your answers as you navigate through the assessment by using the back arrows.
There is a practice question before the assessment so that you can practice answering the questions and using the navigation arrows.
Use the forward arrow at the bottom of this page to begin.

Choose the single best answer:

Sample Question S: When a gas-filled balloon is placed in the freezer so that its temperature decreases, what happens to the volume of the air inside the balloon?

- ○ It decreases.
- ○ It stays the same.

- ○ It increases.

Indicate whether each statement about the gas molecules in the balloon is true or false.

S.1 The gas molecules in the balloon get smaller when it is in the freezer.

- ○ True
- ○ False

S.2 The space between the gas molecules decreases when the balloon is in the freezer.

- ○ True
- ○ False

S.3 The space between the gas molecules increases when the balloon is in the freezer.

- ○ True
- ○ False

S.4 The gas molecules stay the same size when the balloon is in the freezer.

- ○ True
- ○ False

Choose the single best answer:

Question 1: A coin is tossed straight upward. What is/are the force(s) that act on the coin after it has been released and as it travels upward? (Ignore air resistance.)

- ○ No forces
- ○ An upward force only
- ○ A downward force only
- ○ Both an upward and a downward force

Indicate whether each of the following statements about the coin's motion and the forces acting on the coin is true or false.

1.1 The coin slows down as it gets higher.

- ○ True
- ○ False

1.2 Gravity does not act on the coin while it is moving upward.

- ○ True
- ○ False

1.3 The coin slows down because the total force decreases as the coin moves upward.

- ○ True
- ○ False

1.4The total force stays the same as the coin gets higher.

- ○ True
- ○ False

Choose the single best answer.

Question 2: A coin is tossed straight upward. When the coin is at the top of its path, it is momentarily at rest because it has stopped moving up and has not started to fall yet. What force(s) act on the coin when it is at this point? (Ignore air resistance.)

- ○ No forces
- ○ An upward force only
- ○ A downward force only
- ○ Both an upward force and a downward force

Indicate whether each of the following statements about the forces acting on the coin and the coin's motion is true or false.

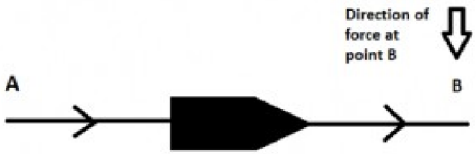
2.1 Gravity does not act on the coin when it is at the top of its path.

- ○ True
- ○ False

2.2 The coin's acceleration is zero when it is at the top of its path.

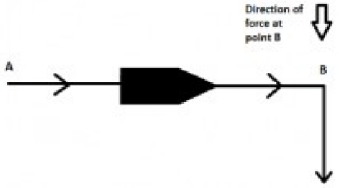
- ○ True
- ○ False

For questions 3 and 4: A rocket is traveling sideways through space at a constant speed from point A to point B as shown. It is far from any planets or anything else that might exert a gravitational force on it.
When the rocket reaches point B, its engines are turned on and left on which creates a force in the direction shown.



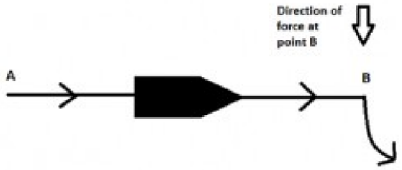
Question 3: Which picture best describes the rocket's path including the part after it passes point B?

a.

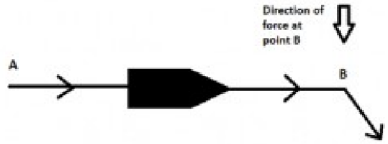


b.

c.



d.



Choose the single best answer.

- ○ Picture a
- ○ Picture b
- ○ Picture c
- ○ Picture d

Indicate whether each of the following statements about the rocket is true or false.

3.1 There is no force pushing the rocket from A to B.

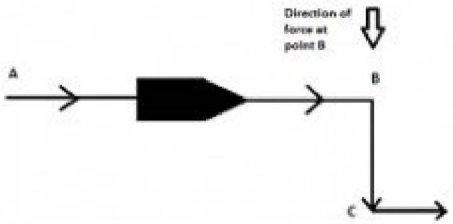
- ◯ True
- ◯ False

3.2 The rocket changes direction because the engine pushes the fuel exhaust into space.

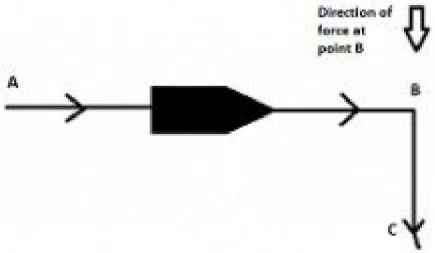
- ◯ True
- ◯ False

Question 4: The same rocket's engines are turned off once the rocket reaches point C. Which picture best describes the rocket's path including the part after it passes point C?

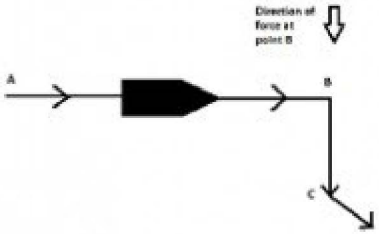
a.



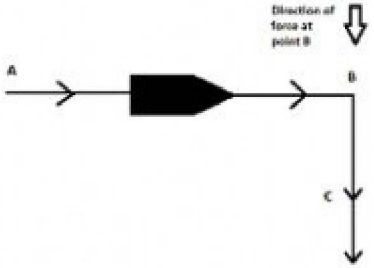
b.



c.



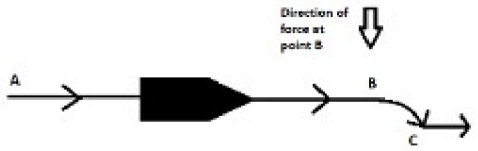
d.



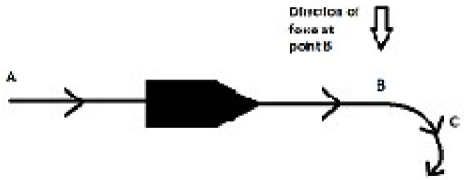
Choose the single best answer.

- ◯ Picture a
- ◯ Picture b
- ◯ Picture c
- ◯ Picture d

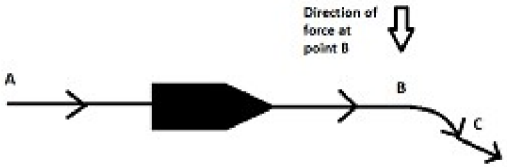
a.



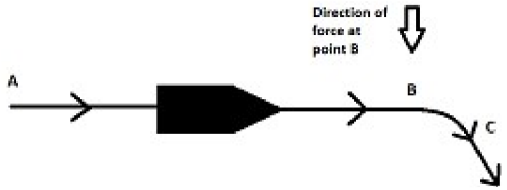
b.



c.



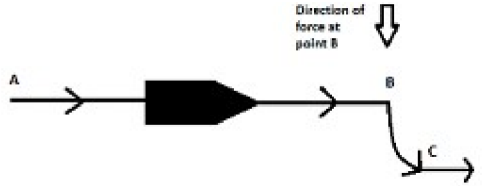
d.



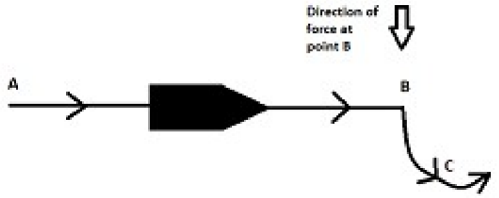
Choose the single best answer.

- ○ Picture a
- ○ Picture b
- ○ Picture c
- ○ Picture d

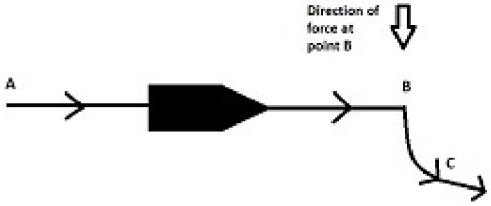
a.



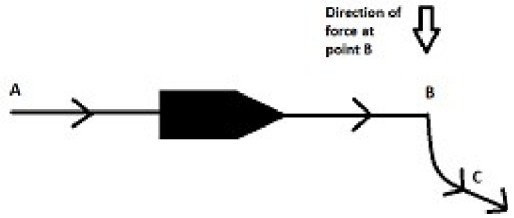
b.



c.



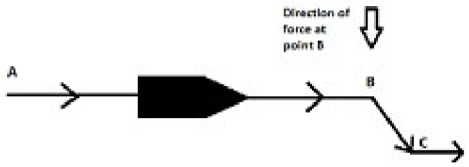
d.



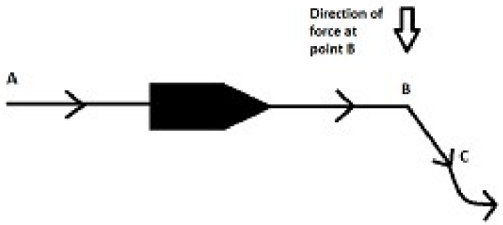
Choose the single best answer.

- ○ Picture a
- ○ Picture b
- ○ Picture c
- ○ Picture d

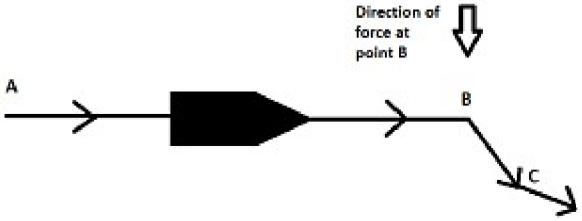
a.



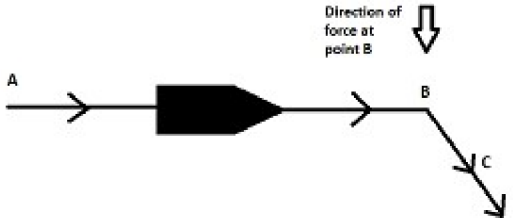
b.



c.



d.



Choose the single best answer.

- ○ Picture a
- ○ Picture b
- ○ Picture c
- ○ Picture d

Indicate whether each of the following statements about the rocket is true or false.

4.1 When the engine is turned off at point C, the rocket keeps moving in the same way because no forces are acting on it.

- ○ True

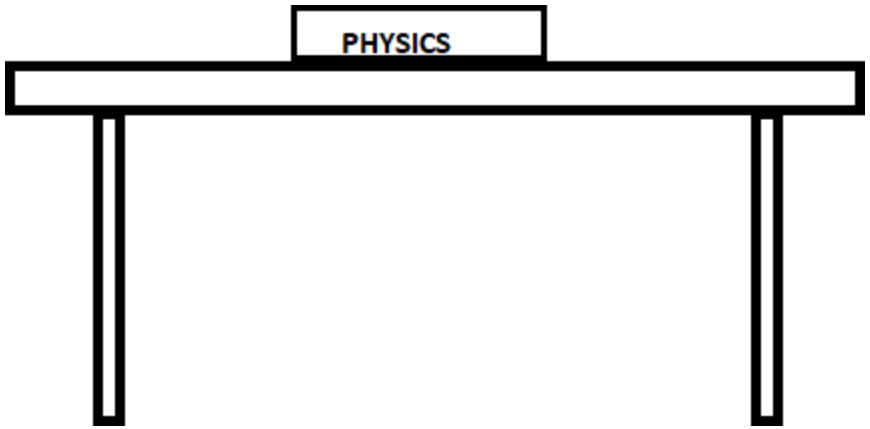
○ False

4.2 After the engine is turned off, the original force makes the rocket move in the direction shown.

○ True

○ False

Question 5: A book is at rest on a table as shown. What force(s) are acting on the book? There is no wind in the room.



Choose the single best answer.

○ No forces are acting on the book

○ An upward force only

- ○ A downward force only
- ○ An upward force and a downward force

Indicate whether each of the following statements about the book is true or false.

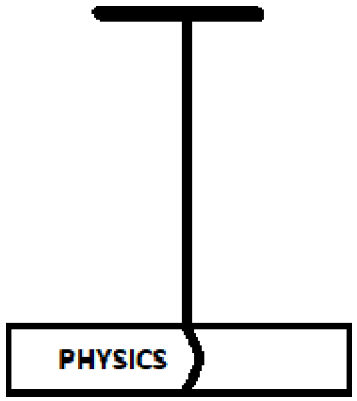
5.2 The table does not exert a force on the book, it simply gets in the way to keep the book from falling.

- ○ True
- ○ False

5.3 There are no horizontal forces acting on the book.

- ○ True
- ○ False

Question 6: A book is at rest hanging from a string as shown. Which forces are acting on the book? There is no wind in the room.



Choose the single best answer.

- ○ No forces are acting on the book.
- ○ An upward force only
- ○ A downward force only
- ○ An upward force and a downward force

Indicate whether each of the following statements about the book is true or false.

6.1 There are no horizontal forces acting on the book.

- ○ True
- ○ False

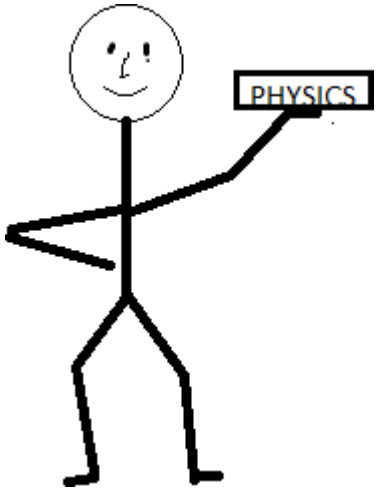
6.2 The book is at rest; therefore, the force of gravity is not pulling it down.

- ○ True
- ○ False

6.3 The string exerts an upward force on the book.

- ○ True
- ○ False

Question 7: A book is at rest on a person's hand as shown. How many forces are acting on the book? There is no wind in the room and the book is horizontal.



Choose the single best answer.

- ○ No forces act on the book.

- ◯ An upward force only.
- ◯ A downward force only
- ◯ An upward force and a downward force

Indicate whether each of the following statements about the book is true or false.

7.1 There are no horizontal forces acting on the book.

- ◯ True
- ◯ False

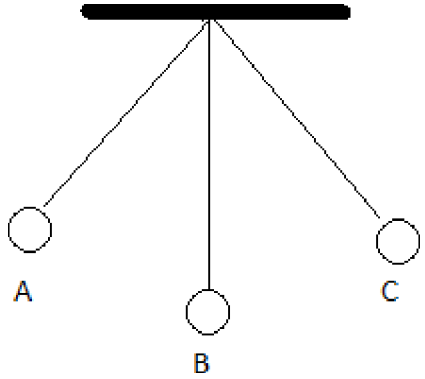
7.2 The book is at rest; therefore the force of gravity is not pulling it down.

- ◯ True
- ◯ False

7.3 The hand exerts an upward force on the book.

- ◯ True
- ◯ False

For questions 8 & 9: A ball is hung from a string attached to the ceiling. A person raises the ball to Point A and releases it so that it swings back and forth. The ball speeds up and slows down as it swings. When the ball is at Points A and C, it is momentarily at rest. When the ball is at Point B, it is going fastest. Assume that there is no air resistance.



Choose the single best answer.

Question 8: How many forces are acting on the ball when it is moving from point B to point A?

- ○ Zero
- ○ One
- ○ Two
- ○ Three

Indicate whether each of the following statements about the ball is true or false.

8.1 The ball is not falling; therefore, the force of gravity is not pulling it down.

- ○ True

○ False

8.2 The ball slows down as it gets closer to point A.

○ True

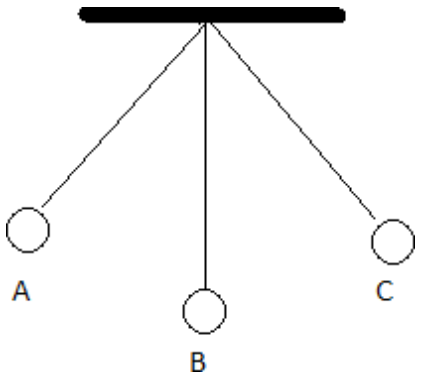
○ False

8.3 There is a force in the direction of the ball's motion and this force gets smaller as the ball slows down.

○ True

○ False

For questions 8 & 9: A ball is hung from a string attached to the ceiling. A person raises the ball to Point A and releases it so that it swings back and forth. The ball speeds up and slows down as it swings. When the ball is at Points A and C, it is momentarily at rest. When the ball is at Point B, it is going fastest. Assume that there is no air resistance.



Choose the single best answer.

Question 9: How many forces are acting on the ball when it is at point B?

- ○ Zero
- ○ One
- ○ Two
- ○ Three

Indicate whether each of the following statements about the ball is true or false.

9.1 Gravity does not pull on the ball when it is at its lowest point.

- ○ True
- ○ False

9.2 The string does not exert a force on the ball.

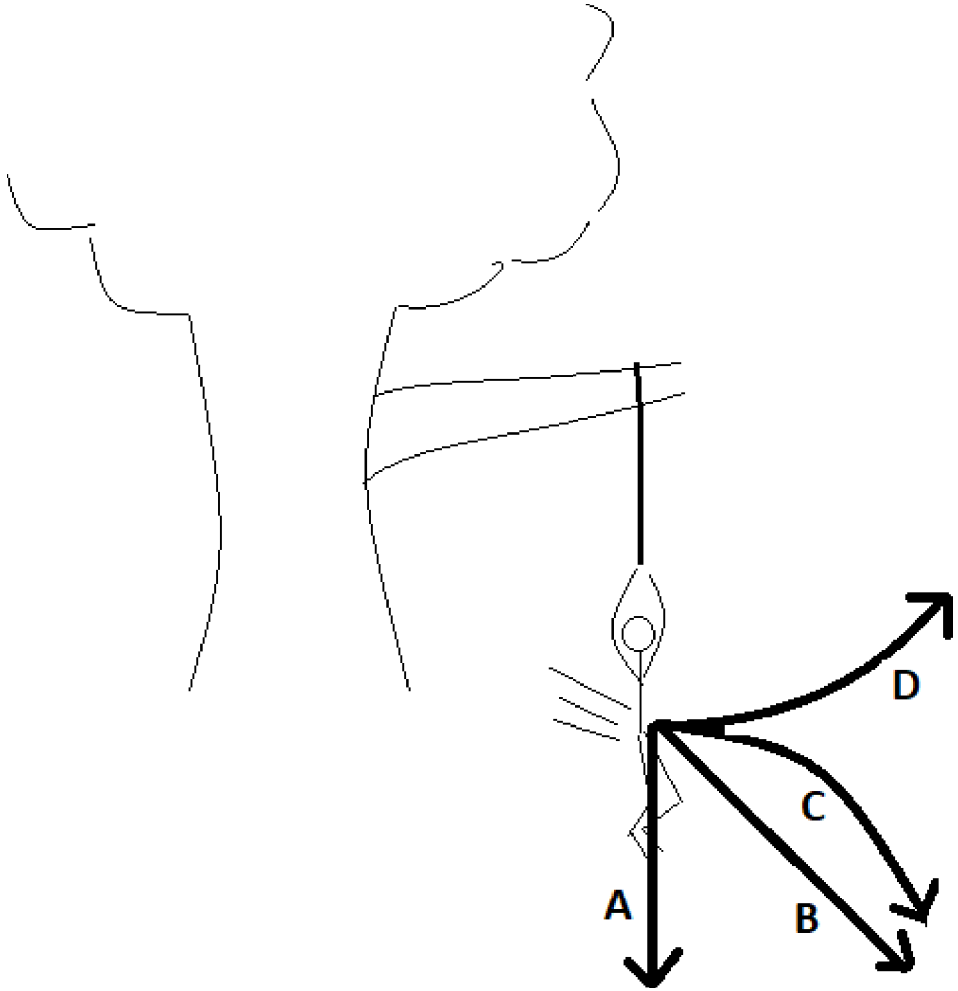
- ○ True
- ○ False

9.3 There is a force pushing the ball in the direction of motion and the force is largest when the ball is at point B.

- ○ True

○ False

Question 10: A person swings from a rope swing above a lake. The person starts high in the tree and lets go of the rope when they are at the bottom of the swing. Which line best shows the path the person takes as they fall into the lake?



Choose the single best answer.

- ○ Path A
- ○ Path B
- ○ Path C
- ○ Path D

Indicate whether each of the following statements about the person is true or false.

10.1 There is a constant force pushing the person forward as they fall.

- ○ True
- ○ False

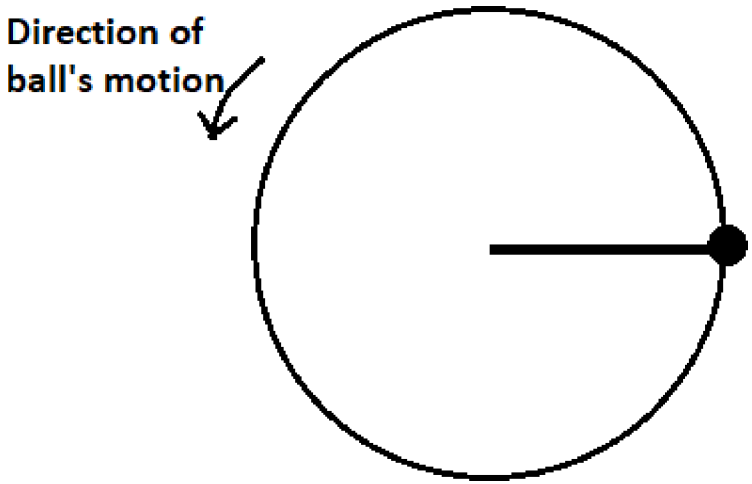
10.2 There is a force pushing the person forward that gets smaller as they fall.

- ○ True
- ○ False

10.3 Gravity pulls the person downward as they fall.

- ○ True
- ○ False

For questions 11 and 12: A ball is tied to the end of a string and being swung along a circular path around a person's head at a constant speed when the string breaks. Imagine you are looking at the ball from directly above. The picture shows what you would see. There is no air resistance.



Question 11: How many forces act on the ball ***before*** the string breaks?

- ○ One
- ○ Two
- ○ Three
- ○ Four

Indicate whether each of the following statements about the ball before the string breaks is true or false.

11.1 There is a constant force that pushes the ball in the direction of motion.

- ○ True
- ○ False

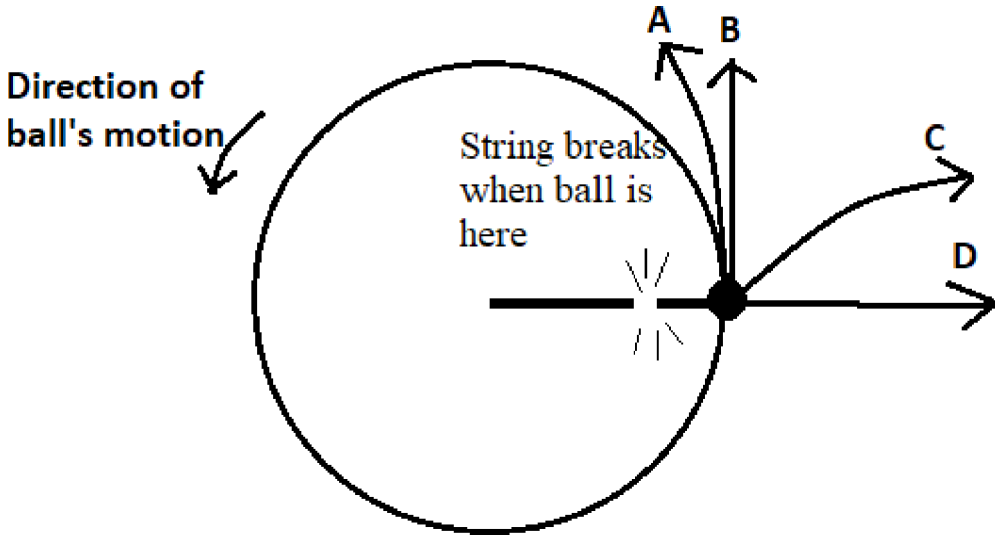
11.2 There is a constant force that pushes the ball away from the center of the circle.

- ○ True
- ○ False

11.3 The ball is pulled toward the center of the circle by the string.

- ○ True
- ○ False

Question 12: Imagine that you are viewing the ball from directly above. The string breaks when the ball is at the position shown. Which path would you see the ball follow after the string breaks?



Choose the single best answer.

- ○ Path A
- ○ Path B
- ○ Path C
- ○ Path D

Indicate whether each of the following statements about the ball is true or false.

12.2 The only force acting on the ball as it falls is gravity.

- ○ True

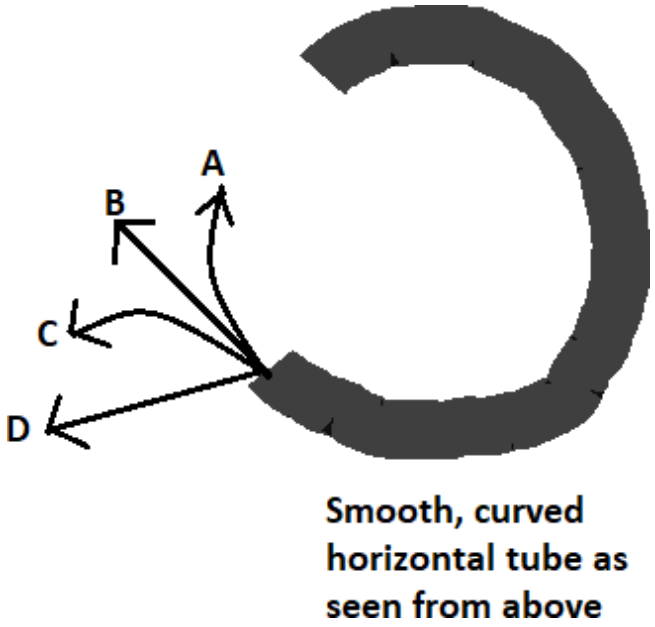
○ False

12.3 There is a force pulling the ball away from the center of the circle as it falls.

○ True

○ False

Question 14: A puck is shot through a smooth, curved, horizontal tube. The tube is attached to the floor. The floor is smooth so that the puck can slide across it without slowing down. The picture shows the tube and puck as seen from above. Which line best describes the path the puck takes when it leaves the tube? (There is no friction or air resistance acting on the puck as it slides across the table.)



Smooth, curved horizontal tube as seen from above

Choose the single best answer.

- ○ Path A
- ○ Path B
- ○ Path C
- ○ Path D

Indicate whether each of the following statements about the puck is true or false.

14.2 There are no horizontal forces acting on the puck as it slides across the floor after leaving the tube.

- ○ True
- ○ False

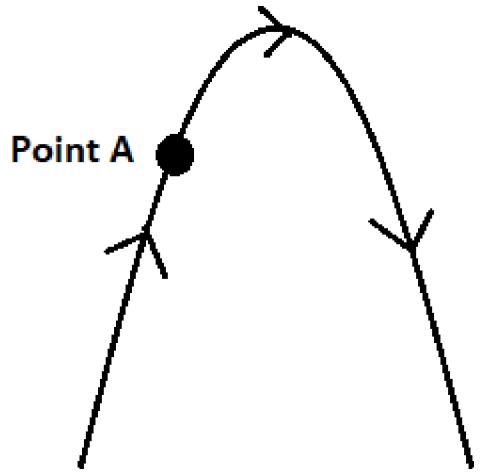
14.3 There is a force pulling the puck toward the outside of the circle when it is in the tube.

- ○ True
- ○ False

14.4 After the puck leaves the tube and is sliding across the floor, there is a force acting in the direction of motion.

- ○ True
- ○ False

Question 15: A juggler throws a ball into the air so that it follows the path shown. How many forces are acting on the ball when it is at point A? Ignore air resistance.



Choose the single best answer.

- ○ Zero
- ○ One
- ○ Two
- ○ Three

Indicate whether each of the statements about the forces acting on the ball when it is at point A is true or false.

15.1 The force of gravity does not act on the ball as it is moving upward.

○ True

○ False

15.2 There is a constant force in the direction of the ball's motion.

○ True

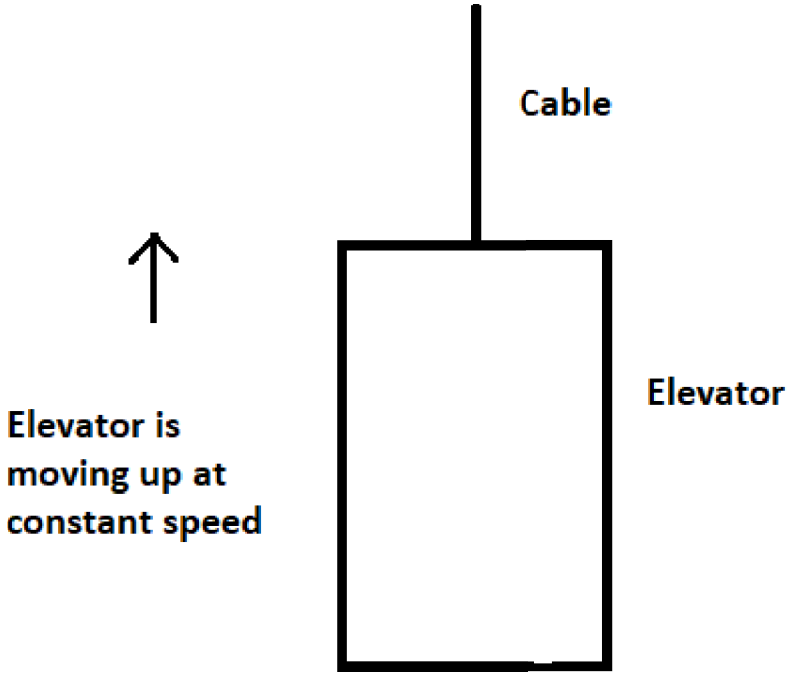
○ False

15.3 There is a force in the direction of the ball's motion and the force gets smaller as the ball approaches the top of its path.

○ True

○ False

Question 16: An elevator that is lifted by a cable as shown is moving upward at a constant speed. What forces are acting on the elevator?



Choose the single best answer

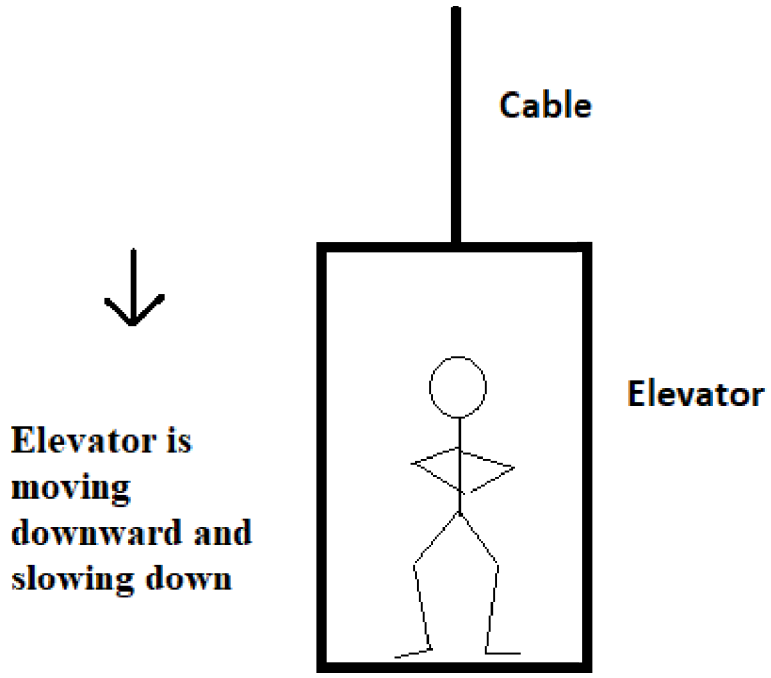
- ○ A single upward force only
- ○ A single downward force only
- ○ A single upward force and a single downward force
- ○ Two upward forces and a single downward force

Indicate whether each of the following statements about the elevator is true or false while it is moving upward at a constant speed.

16.1 The force of gravity does not act on the elevator while it is moving upward.

- ○ True
- ○ False

Question 17: A person is in an elevator that is suspended by a cable and is moving downward and slowing down as shown. What forces are acting on the person?



Choose the single best answer.

- ○ A single downward force only

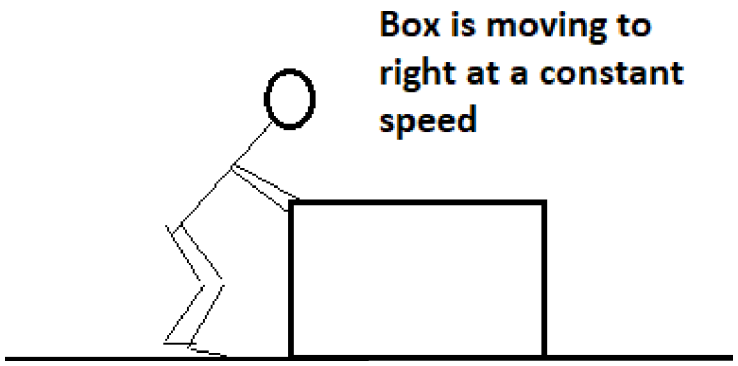
- ○ Two downward forces
- ○ A single upward force and a single downward force
- ○ One upward force and two downward forces

Indicate whether each of the following statements about the person is true or false.

17.1 The person slows down because the downward force gets smaller

- ○ True
- ○ False

Question 18: A person is pushing a box across a horizontal floor so that the box moves at a constant speed to the right. There is friction between the box and the floor. How many forces act on the box while it is being pushed? Ignore air resistance.



Choose the single best answer.

- ○ One
- ○ Two
- ○ Three
- ○ Four

Indicate whether each of the following statements about the box is true or false.

18.1 Neither gravity nor the floor exert a force on the box.

- ○ True
- ○ False

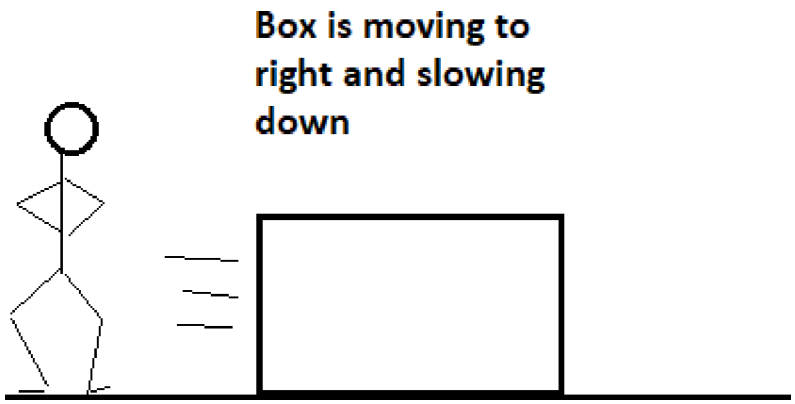
18.2 Gravity exerts a downward force on the box.

- ○ True
- ○ False

18.3 There is a force pushing the box to the right that is bigger than the friction force acting on the box.

- ○ True
- ○ False

Question 19: A box is sitting at rest on a floor. A person walks up to the box, pushes it to the right, and lets it go. The box is sliding to the right across the floor and slowing down. There is friction between the box and the floor. How many forces act on the box while it is sliding to the right and slowing down? Ignore air resistance.



Choose the single best answer.

- ○ One
- ○ Two
- ○ Three
- ○ Four

Indicate whether each of the following statements about the box is true or false.

19.1 The force of the push acts on the box as it is sliding and gets smaller as the box slows down.

- ○ True

- False

19.2 Gravity exerts a downward force on the box while it is sliding.

- True
- False

19.3 Neither gravity nor the floor exert a force on the box while it is sliding.

- True
- False

20. A woman exerts a constant horizontal force on a large box. As a result, the box moves across a horizontal floor at a constant speed "v0".

The constant horizontal force applied by the woman

- has the same magnitude as the weight of the box.
- is greater than the weight of the box.
- has the same magnitude as the total force which resists the weight of the box.
- is greater than the total force which resists the motion of the box.
- is greater than either the weight of the box or the total force which resists its motion

20. A boy throws a steel ball straight up. Consider the motion of the ball only after it has left the boy's hand but before it touches the ground, and assume that forces exerted by the air are negligible. For these conditions, the force(s) acting on the ball is (are)

○ a downward force of gravity along with a steadily decreasing upward force.

○ a steadily decreasing upward force from the moment it leaves the boy's hand until it reaches its highest point; on the way down there is a steadily increasing downward force of gravity as the object gets closer to the earth.

○ an almost constant downward force of gravity along with an upward force that steadily decreases until the ball reaches its highest point; on the way down there is only a constant downward force of gravity.

○ an almost constant downward force of gravity only.

○ none of the above. The ball falls back to the ground because of its natural tendency to rest on the surface of the earth.

Thank you for completing this assessment! Your input is important for making valid interpretations of assessment responses.
Please indicate how you would like to be paid (Venmo, PayPal, GooglePay, or by check) and the email or physical address associated with the account.

Appendix B

Q-Matrix for Final Version of MAFA

Knowledge Item	Reason Item	Knowledge Item Answer	M1	M2	M3	M4	M5	M6
Q1	Q1.1	A, B, C, D	0	0	0	0	0	0
	Q1.2	A, B, C, D	0	0	0	1	0	0
	Q1.3	A	0	0	0	0	0	0
		B, C, D	1	0	0	1	0	0
	Q1.4	A, C	0	0	0	0	0	0
		B, D	1	0	0	0	0	0
Q2	Q2.1	A, B, C, D	1	0	0	1	0	0
	Q2.2	A, B, C, D	0	0	0	0	0	0
Q3	Q3.1	A, B, C, D	1	0	0	0	0	0
	Q3.2	A, B, C, D	0	0	0	0	0	0
Q4	Q4.2	A, C, D	0	0	0	0	0	0
		B	0	0	0	0	1	0
Q5	Q5.2	A, B, C, D	0	0	0	0	0	1
	Q5.3	A, B, C, D	0	0	0	0	0	0
Q6	Q6.1	A, B, C, D	0	0	0	0	0	0
	Q6.2	A, B, C, D	0	0	0	1	0	0
	Q6.3	A, B, C, D	0	0	0	0	0	1
Q7	Q7.1	A, B, C, D	0	0	0	0	0	0
	Q7.2	A, B, C, D	0	0	0	1	0	0
	Q7.3	A, B, C, D	0	0	0	0	0	1
Q8	Q8.1	A, B, C, D	0	0	0	1	0	0
	Q8.2	A, B, C, D	0	0	0	0	0	0
	Q8.3	A	0	0	0	0	0	0
		B, C, D	1	0	1	0	0	0
Q9	Q9.1	A, B, C, D	0	0	0	1	0	0
	Q9.2	A, B, C, D	0	0	0	0	0	1
	Q9.3	A, B, C, D	1	0	1	0	0	0
Q10	Q10.1	A, B, C, D	1	0	0	0	0	0
	Q10.2	A, B, C, D	1	0	1	0	0	0
	Q10.3	A, B, C, D	0	0	0	0	0	0

Knowledge Item	Reason Item	Knowledge Item Answer	M1	M2	M3	M4	M5	M6
Q11	Q11.1	A, B, C, D	1	0	1	0	0	0
	Q11.2	A, B, C, D	0	1	1	0	0	0
	Q11.3	A, B, C, D	0	0	0	0	0	0
Q12	Q12.2	A, C	1	0	0	0	1	0
		B, D	0	0	0	0	0	0
	Q12.3	A, B, C, D	0	1	0	0	0	0
Q14	Q14.2	A, C	0	0	0	1	1	0
		B, D	0	0	0	1	0	0
	Q14.3	A, B, C, D	0	1	0	0	0	0
	Q14.4	A, B, C, D	1	0	0	0	1	0
Q15	Q15.1	A, B, C, D	0	0	0	1	0	0
	Q15.2	A	0	0	0	0	0	0
		B, C, D	1	0	0	0	0	0
	Q15.3	A	0	0	0	0	0	0
		B, C, D	1	0	1	0	0	0
Q16	Q16.1	A, B, C, D	0	0	0	1	0	0
Q17	Q17.1	A, B, C, D	1	0	1	0	0	0
Q18	Q18.1	A, B, C, D	0	0	0	0	0	1
	Q18.2	A, B, C, D	0	0	0	1	0	0
	Q18.3	A, B, C, D	1	0	1	0	0	0
Q19	Q19.1	A, B, C, D	1	0	1	0	0	0
	Q19.2	A, B, C, D	0	0	0	1	0	0
	Q19.3	A, B, C, D	0	0	0	1	0	1

Appendix C

Virginia Tech IRB Approval

Division of Scholarly Integrity and Research Compliance
Institutional Review Board
North End Center, Suite 4120 (MC 0497)
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732
irb@vt.edu
http://www.research.vt.edu/sirc/hrpp

MEMORANDUM

DATE: November 18, 2019

TO: Gary E Skaggs, Mary Norris, Yasuo Miyazaki, David John Kniola, George Glasson

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires October 29, 2024)

PROTOCOL TITLE: Developing an Assessment to Diagnose Physics Misconceptions

IRB NUMBER: **18-917**

Effective November 18, 2019, the Virginia Tech Human Research Protection Program (HRPP) and Institutional Review Board (IRB) determined that this protocol meets the criteria for exemption from IRB review under 45 CFR 46.101(b) category(ies) 2.

Ongoing IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit a new request to the IRB for a determination.

This exempt determination does not apply to any collaborating institution(s). The Virginia Tech HRPP and IRB cannot provide an exemption that overrides the jurisdiction of a local IRB or other institutional mechanism for determining exemptions.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

https://secure.research.vt.edu/external/irb/responsibilities.htm

(Please review responsibilities before beginning your research.)

PROTOCOL INFORMATION:

Determined As:	**Exempt, under 45 CFR 46.101(b) category(ies) 2**
Protocol Determination Date:	**November 21, 2018**

ASSOCIATED FUNDING:

The table on the following page indicates whether grant proposals are related to this protocol, and which of the listed proposals, if any, have been compared to this protocol, if required.

SPECIAL INSTRUCTIONS:

This amendment, submitted October 30, 2019, updates research protocol to extend participant recruitment for phases 3 and 4 to other four-year public institutions of higher education in Virginia. These are: Christopher Newport University, William and Mary, George Mason University, University of Virginia, James Madison University, Radford University, Old Dominion University, Norfolk State University, Virginia Commonwealth University, UVA-Wise, University of Mary Washington, Virginia Military Institute, and Virginia State University. Consent forms were updated to updated IRB name and contact information.

Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this protocol is to cover any other grant proposals, please contact the HRPP office (irb@vt.edu) immediately.

Appendix D

George Mason University IRB Approval

Mason Institutional Review Board <irb@gmu.edu> Fri, Apr 24, 2020 at 8:24 AM
To: Mary Norris <mnorris@vt.edu>

Hi Mary,

You may begin your project since you had already sent us the documents that we requested in January.

If you have any additional questions, please let me know.

Thank you and stay safe,
Kim

Kim Paul, MA
IRB Compliance Specialist
Office of Research Integrity and Assurance
George Mason University
Research Hall, Room 141

Telephone: (703) 993-4208 **Note that I am teleworking, so please email for a more immediate reply**
Fax: (703) 993-9590

Appendix E

Christopher Newport University IRB Approval

From: **Alice Veksler** <alice.veksler@cnu.edu>
Date: Thu, Jan 23, 2020 at 8:34 AM
Subject: Re: Seeking Approval for Research
To: Mary Norris <mnorris@vt.edu>

Hi Mary,
Because you are not affiliated with CNU, the CNU IRB would not generally review your project since you are not conducting work under the supervision of our institution. If the only interaction you have with CNU is requesting faculty to offer the study to their students, IRB review on our end is not needed. If you are looking to post fliers, you should contact student affairs to find out if there are campus rules pertaining to posting things but again, this would be outside the purview of IRB. As long as you have approval from your home institutions' IRB, and your only interaction with CNU is participant recruitment, our IRB will not need to review your project. I appreciate you reaching out to confirm.
With kind regards,
Alice

Alice E. Veksler, Ph.D. | Associate Professor & Chair
| Christopher Newport University | Department of Communication

| Chair, Institutional Review Board (IRB)
| Director, Health Communication Research Lab
| **phone:** 757-594-7461
| **email:** Alice.Veksler@cnu.edu
| **office :** Luter 255A

Appendix F

University of Virginia IRB Approval

From: Blackwood, Bronwyn L (blb2u) <blb2u@virginia.edu>

Date: Thu, Feb 6, 2020 at 10:52 AM

Subject: RE: Seeking Approval for Research

To: Mary Norris <mnorris@vt.edu>

Hi Mary, Robert Jones, the Chair of Physics has said yes. You may contact him to discuss how best to approach students from that angle. I am waiting to hear from the Dean of Students to see if you are allowed to recruit in person on campus (aside from whatever Dr. Jones allows). I'm also waiting to hear from my contact at Wise. Anyway, it's a start. Will get back with you. Bronwyn

Ms Bronwyn L Blackwood

Director, Institutional Review Board for the Social and Behavioral Sciences

Office of the Vice President for Research, University of Virginia

PO Box 800392

Charlottesville, VA 22908-0392

Tel 434-243-2915

Fax 434-924-1992

http://www.virginia.edu/vpr/irb/sbs/

From: Blackwood, Bronwyn L (blb2u) <blb2u@virginia.edu>

Date: Thu, Feb 6, 2020 at 11:34 AM

Subject: RE: Seeking Approval for Research

To: Mary Norris <mnorris@vt.edu>

Hi Mary, Just hear from Mark Clark at UVA Wise. He grants permission as long as your recruiting methods do not involve work on their part. My sense was that you may contact faculty and ask if they would be willing to announce in class, or have you visit the class, and he was okay with your standing in front of the library. I'm still waiting to hear if that is an option for you here at UVA Charlottesville. Best, Bronwyn

Ms Bronwyn L Blackwood

Director, Institutional Review Board for the Social and Behavioral Sciences

Office of the Vice President for Research, University of Virginia

PO Box 800392

Charlottesville, VA 22908-0392

Tel 434-243-2915

Fax 434-924-1992

http://www.virginia.edu/vpr/irb/sbs/

Appendix G

Radford University IRB Approval

From: irb-iacuc <**irb-iacuc@radford.edu**>
Date: Tue, Jan 14, 2020 at 11:02 AM
Subject: RE: Seeking Approval for Participant Recruitment
To: Mary Norris <**mnorris@vt.edu**>
Cc: Lee, Anna Marie <**alee16@radford.edu**>

Good Morning Mary,

Thank you for your note. Please allow me to introduce myself. I am the Research Compliance Manager and newest member of the College of Graduate Studies and Research. I arrived in December and wanted to reach out after seeing your note yesterday. Has your VA Tech IRB protocol been approved? If so and you have noted all of the necessary recruitment and consent documents in your VA Tech approved submission, you ought to be able to move forward. Did someone tell you it was necessary to have Radford IRB approval?

I look forward to hearing from you

Best regards,

Anna Marie

Anna Marie Lee, MHA, CPIA

Research Compliance Manager

Buchanan House

540.831.5290

https://www.radford.edu/content/research-compliance/home.html

From: irb-iacuc <irb-iacuc@radford.edu>

Date: Thu, Jan 16, 2020 at 3:28 PM

Subject: RE: Seeking Approval for Participant Recruitment

To: Mary Norris <mnorris@vt.edu>

Thank you for your note and for the document, Mary. I am forwarding to our IRB persons and will let you know if there is anything else needed.

Best regards,

Anna Marie

Anna Marie Lee, MHA, CPIA

Research Compliance Manager

Buchanan House

540.831.5290

https://www.radford.edu/content/research-compliance/home.html

Appendix H

James Madison University IRB Approval

From: **Morgan, Cindy - morgancs <morgancs@jmu.edu>**
Date: Mon, Dec 2, 2019 at 2:24 PM
Subject: IRB Notice of Exemption from James Madison University
To: **mnorris@vt.edu <mnorris@vt.edu>**

Dear Mary,

I want to let you know that your IRB protocol entitled, "***Developing an Assessment to Diagnose Physics Misconceptions***" has been approved for you to begin your study. The exemption notice memo is attached to this email. Your protocol has been assigned No. 20-0012 for tracking purposes. Thank you again for working with us to get your protocol approved.

If you have any questions, please do not hesitate to contact me.

Best Wishes,

Cindy

Cindy Morgan

IRB Coordinator

Office of Research Integrity - James Madison University

Engineering/Geosciences Bldg., Room 3152

MSC 5738

Harrisonburg, VA 22807

morgancs@jmu.edu